STATIN-ASSOCIATED ADVERSE EVENTS PREDICTION AND DRUG-DRUG INTERACTIONS FOR CARDIOVASCULAR DISEASE PATIENTS FROM RETROSPECTIVE CLAIMS DATA

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

JIN WANG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISOR: TERRENCE J ADAM, RPH, MD, PHD

CO-ADVISOR: CHIH-LIN CHI, PHD, MBA

AUGUST 2019

Acknowledgements

I would like to express my sincere gratitude to both Dr. Terry Adam, my advisor, for his patience, motivation, immense knowledge and his valuable guidance helped me in all the time of research, and Dr. Chih-Lin Chi, my co-advisor, for providing this great opportunity to work on this meaningful project and his continuous encouragement and support throughout my PhD study. Their patience and support helped me overcome numerous obstacles I have been facing through my research and successfully complete my dissertation.

I would especially like to thank my thesis committee members, Dr. Wendy St. Peter and Dr. Angie Carlson, for their insightful comments, encouragement, and the hard questions that incented me to widen my research from various perspectives.

I would also like to take this opportunity to express gratitude to all of the department faculty members and department staff for their help and support, especially Dr. David Pieczkiewicz, Dr. Thomas Clancy, Dr. Rui Zhang, Jessica Whitcomb-Trance, and Melissa Malikowski.

Special thanks to the OptumLabs® staff, Greta Bagshaw, Jamie Tucker, and Jessica Chang, for their kind advice and support that helped me get acquainted with the OptumLabs Data Warehouse.

Also, I would like to thank the other students and classmates, Zhen Hu, Era Kim, Shauna Overgaard, and Wonsuk Oh, for the wonderful time we spent together.

Finally, I would like to thank my parents, my husband, and my daughter for their unconditional love and support.

Dedication

To my parents, Jianqin Li and Jing Wang, my husband, Huanan Zhang,

and my daughter Chloe M. Zhang

Abstract

Statins are commonly used to lower cholesterol levels for cardiovascular disease (CVD) patients in the primary and secondary prevention of acute events. 26% of American adults over age 40 used statins in 2012 and an estimated 26.4 million U.S. adults could benefit from statin use. Although statins are generally well tolerated and show a relatively good safety profile, concerns have been raised regarding statin associated adverse events (AEs) especially muscle related events, leading to medication non-adherence and discontinuation. Besides, AEs are often caused by potential drug-drug interactions (DDIs) which are responsible for up to 2.8% of hospital admissions[1]. Among CVD patients, combination therapy of statins and other medications is highly likely, which results in altered absorption, distribution, metabolism, or excretion of statins and thus causes adverse events. Traditional AE management approaches may include a statin therapy holiday, lower statin dosage, an alternative statin agent, or non-statin cholesterol-lowering therapy. Currently, there are no tools to effectively predict and reduce the risk of AEs prior to statin therapy initiation. In addition, no population-based studies have focused on a specific statin and a specific interacting drug and differentiated their risks among different study time periods.

In this study, we investigated the effect of combination therapy of simvastatin and several pre-defined high risk interacting drugs, which belong to cytochrome P450 (CYP) 3A4 and/or organic anion transporting polypeptide (OATP) inhibitors, in CVD patients who used simvastatin for secondary

prevention. This could provide some evidence and recommendations for selected interacting drugs used in CVD patients. In addition, we aimed to build a model to predict statin-associated AEs that may reduce the risk of statin associated adverse events and the rate of statin therapy cessation. Several machine learning methods were applied, such as generalized linear model (GLM), support vector machine (SVM), decision tree, random forest, and artificial neural network (ANN). Models were developed and compared for their performance. The best model was selected based on the best performance.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Chapter 1 provides an overview of current statin use, including drug class review, current use of statins in the clinical setting, problems and possible solutions for statins adherence issues, and drug-drug interactions (DDIs) review. The significance and the specific aims of this study will also be discussed.

1.1 Overview of current statin use

1.1.1 Brief drug class review

Statins are a class of drug that inhibit 3-hydroxy-3-methyl-glutaryl-coenzyme A (HMG-CoA) reductase, the pivotal rate-controlling enzyme in the production of cholesterol. Seven statin agents currently available in the United States (U.S.) were included in this project.

Table 1- 1 Statin agents that are approved by Food and Drug Administration

Generic Name of Statins	Brand Name	Year of FDA Approval
Lovastatin	Mevacor®, Altoprev®	1987
Pravastatin	Pravachol®	1991
Simvastatin	Zocor®	1991
Fluvastatin	Lescol®, Lescol® XL	1993
Atorvastatin	Lipitor®	1996
Rosuvastatin	Crestor®	2003
Pitavastatin	Livalo®	2009

FDA = Food and Drug Administration

Table 1-1 shows all seven statin agents that are approved for use by the Food and Drug Administration (FDA). Lovastatin was the first statin introduced in the U.S. Pitavastatin, which is known as Livalo®. Pitavastatin is the latest statin to be introduced to the market. Cerivastatin was not included because it was withdrawn from the market in 2001, due to its high risk of fatal rhabdomyolysis events.

1.1.2 Current and emerging statin use for cardiovascular disease patients

High blood cholesterol is an important risk factor contributing to the development of cardiovascular disease (CVD), the leading cause of death worldwide. Reducing the level of cholesterol can help to reduce the chance of developing CVD as well as to prevent the recurrence of acute events in those with known disease[2,3]. Statins are highly effective drugs for lowering low-density lipoprotein (LDL) cholesterol, non-high-density lipoprotein cholesterol, and apolipoprotein B levels in plasma, all of which are key contributors to CVD [4,5]. Extensive clinical trials and studies support beneficial effects of this class of drugs for the secondary prevention and treatment of atherosclerotic CVD[6] in patients who are at very high risk of developing CVD or have had heart diseases such as myocardial infarction and stroke. Prior clinical studies[5,7] also suggested that statins may be beneficial for the primary prevention of coronary heart disease in patients without a history of CVD. Although other non-statin therapies are alternatives to reduce cholesterol, most of them do not have a comparable atherosclerotic cardiovascular event and mortality reduction as compared with

statins. Proprotein convertase subtilisin/kexin type 9 serine protease (PCSK9) inhibitors are a new type of cholesterol drug. A new study [8] published in 2018 showed that risk of major adverse CVD events was lower among those who receive PCSK9 added to statin therapy than among those who treat with statin alone. However, PCSK9 inhibitors are newer and have less long-term safety data. In addition, they are very expensive compared to statins. Thus, the number of patients who received PCSK9 is much smaller than statin users.

A study from the U.S. Department of Health and Human Services showed that during the 2007–2010 time period, approximately 47% of Americans who were more than 65 years old took statins or other cholesterol-lowering drugs [9]. This usage rate of cholesterol-lowering drugs has increased approximately 7-fold since 1988–1994, due in part to the introduction and acceptance of statin drugs.

1.1.3 Problems and previous solutions for therapy with statins

Although statins are generally safe for the majority of individuals, it has been reported that more than 50% of patients discontinued statin medication within 1 year after treatment initiation [10]. Among the various reasons that lead to statin discontinuation or statin intolerance, adverse events (AEs) is the primary reason for such discontinuation [11,12]. Statin-associated AEs[13–15] include minor side effects such as muscle symptoms, digestive problems, dizziness, and some rare but clinically important events such as myopathy[16,17], rhabdomyolysis [18], liver events [17,19–21], acute kidney injury[17,22,23], drug poisoning events[24], and increased risks for hyperglycemia and cognitive effects [25,26]. One study[27] showed that approximately 17% of statin users had an adverse event during their study

period (from 2000 to 2008). The consequence is that important clinical and treatment benefits of statin use for lowering plasma cholesterol levels (and therefore risk of primary and secondary cardiovascular events) are frequently lost due to discontinuance of statin treatment following occurrence of AEs, leading to increased cholesterol levels and increased cardiovascular events[28].

Not every statin user will have AEs, but some patients may have a greater risk of developing statin AEs than others. Patient-related risk factors include older patients, small body frame, a history of specific diseases (myopathy, creatine kinase elevation, muscular symptoms, liver, and kidney), alcohol use, grapefruit juice consumption (>1 quart/day), major surgery in the perioperative period [29] female patients [11], and excessive physical activity [30]. Treatment-related risk factors include high-dose statin therapy and combination therapy of statins and other medications which are either substrates or inhibitors of cytochrome P450 (CYP) 3A4) [31].

Typical strategies to manage and control statin-related AEs include reducing statin dosage, using an alternative statin agent or non-statin cholesterol-lowering therapy, switching to nondaily dosing statin regimen, using a statin holiday, or dietary intervention[32–34]. In practical settings, these traditional approaches control AEs after the AE occurrence, leading to statin treatment discontinuation. This could be problematic. For example, switching to a lower statin intensity can reduce AE risk, but may also lessen the reduction in atherosclerotic cardiovascular disease risk [34]. Switching to non-statin therapy is problematic as well, given that many non-statins have worse AE profiles and less

effective than statins at lowering LDL cholesterol[35]. Proactive prevention of statin adverse events has not been well studied. Individuals may have different responses to individual statins at a specific dosage. So when physicians prescribe statins, the decision should not only depend on the cholesterol levels but also the patients' characteristics.

1.1.4 Drug interactions

Statin interactions, which include statin-food and statin-drug interactions, are a common cause of AEs. Most of the statins are metabolized by CYP450 enzymes. Simvastatin, atorvastatin, and lovastatin are metabolized by CYP3A4. Fluvastatin and rosuvastatin (to a lesser extent) are substrates for CYP2C9. Pravastatin and pitavastatin are minimally metabolized by CYP450 enzymes. In addition, the organic anion transporting polypeptide (OATP) 1B1 is known to transport all statins from plasma into hepatocytes for metabolism and elimination[36–42]. Hence, food or drugs that potentially are substrates for or inhibit CYP3A4 enzyme or/and inhibit OATP transporters may increase the plasma concentration of statins and thus increase the risk of adverse events.

The most common statin-food interaction, especially for simvastatin, lovastatin, and atorvastatin, is grapefruit juice which contains an organic chemical compound -furanocoumarins that can inhibit the CYP3A4 enzyme and increase statin levels in plasma[43]. Grapefruit also affects OATP transporters[44].

Statin-drug interactions are another type of interaction. Many CVD patients may need statins in combination with other therapies, especially in those who have multiple comorbidities, and those at high CVD risk who cannot achieve

optimal therapeutic benefits from statin monotherapy[45–47]. In one study, more than 20% of statin users were detected with potential statin DDIs[48]. In this study, several interacting medications with the highest DDI risk were included. These medications were selected by clinical expert review after carefully screening available DDIs resources including drugs.com, Lexicomp, Epocrates, and one published literature[49]. Table 1-2 shows the mechanisms of interaction that these medications have with simvastatin, lovastatin, and atorvastatin. Antibiotics such as clarithromycin, erythromycin, and telithromycin are strong inhibitors of the CYP3A4 and OATP1B1. Studies and case reports[50–54] have shown that coprescription of a statin agent with clarithromycin, erythromycin, or telithromycin are associated with a higher risk of rhabdomyolysis, acute kidney injury, and/or all-cause mortality. Antifungals medications such as itraconazole, ketoconazole, and posaconazole are potent CYP3A4 inhibitors. Co-administration of those medications with statins could induce rhabdomyolysis and renal injury [55–57]. Nefazodone[58] (antidepressant drug), boceprevir[59] (protease inhibitor for hepatitis), and danazol[60] (androgenic hormone) are also potent CYP3A4 inhibitors. Many case reports showed patients who were treated with those drugs while on concomitant statins could possibly develop an increased risk of rhabdomyolysis or myopathy [59,61–64]. Cyclosporine[40,65] (immunosuppressant) and cobicistat[66] (pharmacokinetic enhancer) are both CYP3A4 and OATP1B1 inhibitors. Therefore, coadminstration of those with statins is contraindicated because of the high risk of rhabdomyolysis or kidney injury [67–69]. Gemfibrozil[40], a fibric acid agent, is a well-known OATP inhibitor that does not alter CYP3A4 activity. The

combination therapy of gemfibrozil and statins is also considered a contraindicated treatment[70–72].

Table 1- 2 The inhibitors of CYP3A4 or OATP1B1

Drugs	CYP3A4 inhibitor	OATP1B1 inhibitor
Antibiotics (clarithromycin, erythromycin, and telithromycin)	√	√
Antifungals (itraconazole, ketoconazole, and posaconazole)	√	X
Cyclosporine	√	√
Cobicistat	√	√
Nefazodone	√	X
Boceprevir	√	X
Danazol	√	X
Gemfibrozil	X	√

Note: CYP = cytochrome P450;

OATP = organic anion transporting polypeptide

1.2 Significance

This dissertation addresses an important issue - statin associated adverse events in CVD population- which were used to study an important problem of drug therapy optimization. We focused on general statin AEs associated AEs as well as those caused by combination therapy of simvastatin and several pre-defined high risk interacting drugs (CYP3A4 and/or OATP inhibitors).

We utilized existing observational data to evaluate statin AEs. Randomized controlled trials often have limited generalizability because of rigorous inclusion and exclusion criteria for selecting patients. The observational dataset, on the other hand, allows a pragmatic research approach, including patients who have more complicated situations than those in clinical trials. Medication and clinical strategies are not controlled by the researchers' experimental design and can reflect the actual clinical practice patterns without restricted conditions and thus, facilitate high generalizability.

Healthcare machine learning methods were integrated to manage and analyze the large volume of population-based data. The current reactive approach to manage statin AEs is problematic because it only initiated after AEs have occurred. The proactive strategy for statin prescription decision lacks a working approach outside of a trial and error prescribing approach. Few previous studies focused on the prediction of statin associated AEs. This dissertation focused on this aspect by developing predictive model, which is capable of integrating complex patient characteristics as predictors, for predicting statin associated adverse events. The results from big data analysis can provide additional information and support along with experts' personal experience for medical decision making. This model is the first step toward of overcoming the clinical challenges on statin associated AEs reduction and will be adapted to develop a decision-support system - the Personalized Statin Treatment Plan platform, to support proactive statin prescription decisions. This use case can

also be expanded to other medication management issues to reduce AEs and improve the medication adherence.

1.3 Specific Aims

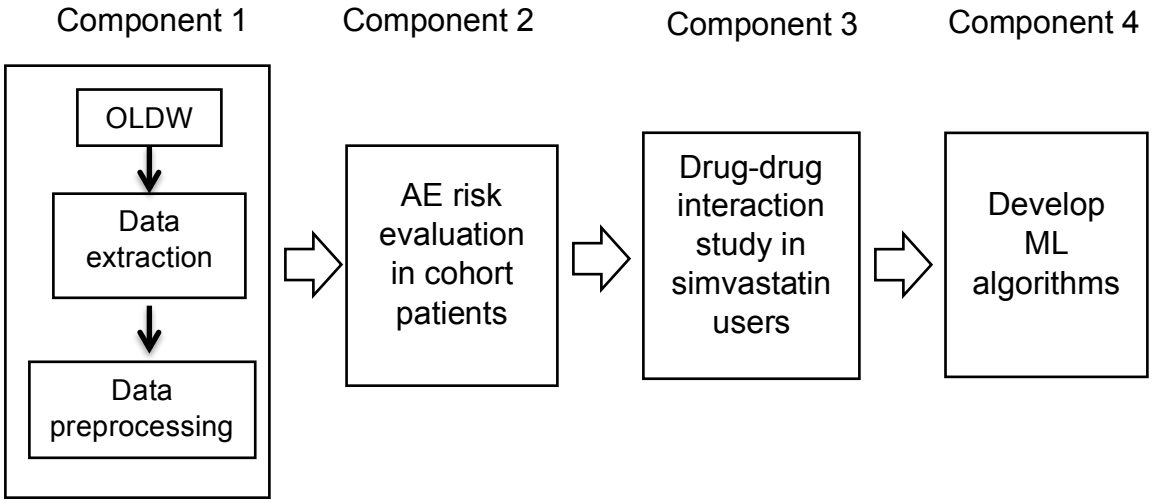
This dissertation hypothesizes that statins combined with medications that have known DDIs with statins would cause more adverse events and more intense medical follow-up. In addition, the development of an adverse events prediction model could identify the risk of adverse events with statins and interacting medications and can better assist clinicians to estimate the risk of adverse events. The objectives of this dissertation are to investigate the adverse events among statin users in CVD patients in order to provide clinical evidence of AEs risk comparison, estimation, and management, as well as to develop a predictive model that can be used in the future to aid clinicians in selecting an optimal statin treatment plan for individual patients to minimize the risk of adverse events. To address these objectives, three specific aims were proposed:

1) To evaluate the risk of adverse events in CVD patients who initiated statin agents for secondary prevention of CVD.
2) To evaluate the effect of combination use of simvastatin with known high-risk drugs which are inhibitors of CYP3A4 and/or OATP transport in CVD patients.
3) To develop a predictive model using machine learning methods for the occurrence of statin associated adverse events in CVD patients.

1.4 Thesis outline

The entire project consists of four components (summarized in Figure 1-1).

The following chapters will introduce these components.



Note: OLDW = OptumLabs® Data Warehouse

ML = machine learning

Figure 1- 1 Project pipeline

Chapter 2 Background

In this chapter, background knowledge including secondary data analysis, data source, machine learning algorithms, cross validation, and model performance metrics will be introduced.

2.1 Emerging Models for Secondary Data Analysis

The use of secondary data analysis (SDA) is an important resource for population-based healthcare research. This, in part, can be attributed to improvements in access to large scale, complex datasets previously described as “big data”.

A number of large databases are available to researchers today. These include public databases such as the Medicare Claims Public Use files (588,415 lives annually)[73] , the Centers for Disease Controls National Center for Health Statistics (42,000 households per year)[74] and the Agency for Healthcare Research and Quality Healthcare Cost and Utilization Project (7 million hospital stays annually)[75]. These databases contain a mix of insurance claims data as well as survey and census data. Other databases may include grant funded projects such as the Clinical Translational Science Institute (62 medical institutions in 32 states)[76], the Patient Centered Outcomes Research Institute (47 million lives of electronic health record and patient reported outcomes data)[77], and the Healthcare Cost Institute (40 million lives annually of insurance claims data) [78]. Finally, with passage of the Health Information Technology for Economic and Clinical Health Act in 2009, health systems have created large scale data repositories of clinical and administrative data from the

implementation of electronic health records. These data repositories, fed by multiple hospitals within an integrated health system, can contain data on millions of patient records.

2.2 Data Source

An emerging research model is the development of partnerships, which make available "primary" use of data for SDA. A pioneer in this new model is OptumLabs. OptumLabs is an open, collaborative research and innovation center founded in 2013 as a partnership between Optum and Mayo Clinic with its core linked data assets in the OptumLabs Data Warehouse (OLDW). It comprises more than 20 years of data and approximately 200 million patient records in this database with more new patients and patient information being added and updated periodically providing additional insights. The database contains de-identified, longitudinal health information on enrollees and patients, representing a diverse mixture of ages, ethnicities and geographical regions across the United States. The claims data in OLDW includes medical and pharmacy claims, laboratory results and enrollment records for commercial and Medicare Advantage enrollees. The EHR-derived data includes a subset of EHR data that has been normalized and standardized into a single database[79].

The administrative claims data includes medical claims, pharmacy claims (both pharmacy and pharmacy Part D data), and enrollment information which includes cost data (patient and health plan paid amount). Administrative data is linkable to health risk assessment, SES information, lab test results (focus on

serum, urine, and blood-based labs), and supplemental oncology data included in OLDW.

The medical claims data are collected from all health care sites (e.g., inpatient hospital, outpatient hospital, emergency room (ER), physician's office, and surgery center) for services including specialty, preventive and office-based treatments. It has one line of data for each service claim and includes data such as date of service, diagnosis codes, procedure codes, site of service codes, present on admission (POA) codes, patient and health plan paid amounts, and provider specialty codes. While the medical claim data is updated monthly, researcher typically allow 4-6 months to account for claim adjudication.

The pharmacy claims data is comprised of outpatient prescription fills. They are submitted by pharmacies, including retail, mail order, hospital discharge, and specialty pharmacies. Pharmacy Part D claims are stored in another table which only includes prescription claims for Medicare Part D plan enrollees. The pharmacy claims data includes drug related information (e.g., medication name, dosage form, drug strength, fill date, and days of supply). While pharmacy data is updated monthly, researchers typically allow 8 weeks to account for adjudication.

Using administrative data for healthcare research presents a number of benefits.1) It contains a large number of patients that can provide the opportunity and possibility to study low incidence events. 2) Its ready availability provides the possibilities of obtaining reliable large sample size of data at low cost compared with traditional censuses and questionnaires. 3) Information included in administrative data are not easily obtained and reported by individuals or

elsewhere. Despite advantages, several limitations also need to be considered. 1) This dataset does not represent the entire US population. For example, it does not include administrative claims for Medicaid plans. Clinical activities for patients who have Medicaid plans are captured in clinical data. 2) Clerical errors may exist in claims data including incorrect or missing data. 3) Diagnosis and procedure codes cannot reliably infer disease severity. 4) These data do not capture all the medical events. Claims-based reporting mainly exists to ensure that reimbursement is obtained by providers for medical services. Some events may not be tied to reimbursement. For example, a patient calling their provider, stopping therapy, not seeking medical attention may not be captured reliably. 5) Only prescribed medications covered under the benefit plan are included. Over-the counter (OTC) medications are not reliably covered and thus many patients may have missing data in terms of OTC medications. In addition, medications that patients choose to pay out-of-pocket, herbal therapies or supplements are generally not included in claims data. This can potentially limit research accuracy. 6) Drug information is extracted from prescription records, but we cannot guarantee patients actually take the prescribed drugs. 7) The lab results in OLDW claims data are incomplete and underrepresented in the data set. While this varies by test, lab results are available for approximately 40% of patients with a laboratory test. However, the presence of a lab result for an individual does not mean that all of their lab results are included. These data only contain outpatient laboratory values, with the most common sample types being serum, urine, and

blood-based samples. Also, results for lab tests processed outside of certain clinical laboratories are not available in the database.

2.3 Machine learning algorithms

Machine learning methods provide powerful approaches to discover hidden patterns from a large amount of data and are primarily used for prediction and exploratory studies. Supervised learning and unsupervised learning are two main types of machine learning algorithms. They are distinguished by whether training data has known output variables (also known as results, labels, or outcomes). For example, to predict whether a patient has hypertension, patient traits (e.g., age, gender, race, medical history, and lab tests) are referred to as input variables/features/predictors and disease information (whether a patient has hypertension) is referred to as output variable. Models. Supervised learning is applied when training data has both input and output variable. It trains a model on these known input and output variables and predicts future outputs on new patients. If only input variables are available, then unsupervised learning methods, which finds hidden patterns in the input data, should be applied. Broadly speaking, supervised learning includes categories of classification and regression, while unsupervised learning includes categories of association rule learning and clustering analysis. For this project, supervised learning methods were applied since the outcomes variable (whether patients had AEs) was available for each individual in our training data.

Five classic supervised learning classification models were initially investigated, including generalized linear models (GLMs), support vector

machine (SVM), decision tree, random forest, and artificial neural networks (ANN).

2.3.1 GLMs

GLMs[80] represent a broad class of regression models including but not limited to linear regression, logistic regression, and Poisson regression. GLMs allow researchers to generalize the linear regression approach to accommodate many types of response variables (e.g., continuous, binary, proportions, count, and positive count data).

In a GLM,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

y_i is the response variable which is modeled by a linear function of explanatory variables $x_j (j = 1, \ldots p)$ plus an error term. It is made up of a linear predictor: $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$, and two functions: 1) a link (mean) function that specifies how the expected value of response variable, $E(Y_i) = \mu_i$, relates to the linear predictor: $g(\mu_i) = \eta_i$, and 2) a variance function that specifies the associations between the variance and the mean: $\mathrm{var}(\mathrm{Y_i}) = \phi \mathrm{V}(\mu)$ where the dispersion parameter ϕ is a constant.

Some common regression model specifications include 1) linear model containing an intercept and linear term for each predictor, 2) interactions model containing an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms), 3) pure-quadratic model contains an intercept term and linear and squared terms for each predictor, and 4) quadratic model containing an intercept term, linear and squared terms for each predictor,

and all products of pairs of distinct predictors. Table 2-1 shows some examples of these models.

Table 2- 1 Model functions of different model specifications

Model specification	**Model functions (Assume has only two predictors $\mathbf{x_1}$ and $\mathbf{x_2}$)**
Linear	$y \sim b_0 + b_1x_1 + b_2x_2$
Interaction	$y \sim b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$
Pure-quadratic	$y \sim b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2$
Quadratic	$y \sim b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2$

The assumptions of GLM include: 1) the response variables are independently distributed; 2) linear relationship between the transformed response in terms of the link function and the explanatory variables; 3) errors need to be independent but not normally distributed.

In GLM, response variables do not need to be transformed to have a normal distribution. The choice of link function is separate from the choice of distribution. Table 2-2 shows the choices of link functions for difference data types. Table 2-3 shows the common link functions.

Table 2- 2 Common data types with model distribution, variance function, and link types

Data type	**Model distribution**	**Variance function, $\mathbf{Var(\mu_i)}$**	**Canonical Link**	**Other Links can be used**
Continuous data	Normal	1	Identity	Log, inverse

Binary data	Bernoulli	$\mu_i(1-\mu_i)$	Logit	Probit, log
Count data	Poisson	μ_i	Log	Identity, square root
Positive counts	Gamma	μ^2	Inverse	Log, identity

Note: μ_i is the expected value of the response variable.

Table 2- 3 GLMs common link functions

Link type	**Link functions, $g(\mu_i)$**
Identity link	μ_i
Logit link	$\ln\left(\frac{\mu_i}{1-\mu_i}\right)$
Log link	$\ln(\mu_i)$
Inverse link	$\frac{1}{\mu_i}$
Square root link	$\sqrt{\mu_i}$
Probit link	$\phi^{-1}(\mu_i)$

Note: μ_i is the expected value of the response variable.

ϕ is the cumulative distribution function (CDF) of the standard-normal distribution.

2.3.2 Support Vector Machine

SVM[81] is one of the top-ranked algorithms that can be used for classification and regression analysis with good generalization performance and good ability to solve a wide range of problems. It performs classification work by drawing a separating line which is known as a hyperplane[82]. The objective of SVM is to find a hyperplane in a N-dimensional (N is the number of features),

given labeled training data, that distinctly categorizes the data points into two classes by maximizing the margin (decision boundary) around the hyperplane. Generalization error improves when the margin is larger. Data points that are closest to the hyperplane are called support vectors. For example, in Figure 2-1, the best hyperplane is the one whose distance to the nearest element of each class (blue circle or red square) is the largest. Data points that located on either side of the hyperplane are identified as different classes. In 2-dimensional space, a hyperplane is a line. In 3-dimensional space, a hyperplane is a plane. As the number of features increase, it becomes hard to imagine the hyperplane.

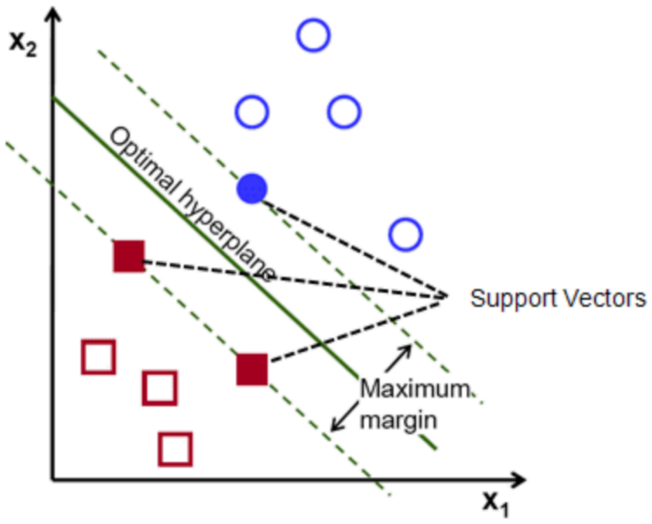


Figure 2- 1 SVM maximum margin and optimal hyperplane

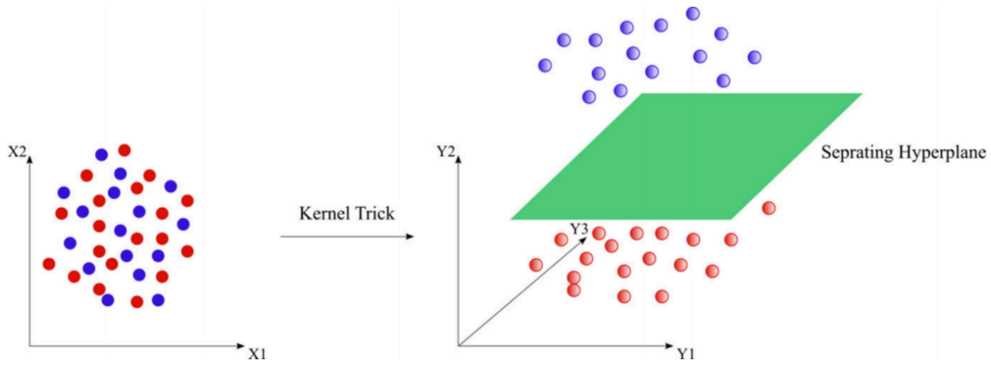


Figure 2- 2 Kernel function for non-linearly separable data points

However, sometimes data points are not linearly separable in the original dimensional space. A kernel function is a method that uses a linear classification method to solve a non-linear problem. The kernel function finds a linear decision boundary by mapping the original non-linear data points into a higher dimensional space. In Figure 2-2, it shows an example that the two classes (blue dots and red dots cannot be separated linearly in the original 2-dimensional space. However, they can be separated by a plane when those data points are projected in a higher dimension (a 3-dimensional space in this example). Kernel functions can be viewed as similarity functions that compute the similarity score (a dot product) of two vectors x_i and x_j in a higher dimensional space. The common kernel functions include linear kernel, radial basis function (RBF) kernel, and polynomial kernel. Table 2-4 shows the formula for different types of kernels. The linear kernel works when data points are linearly separable. The RBF and polynomial kernels apply when data points are not linearly separable.

Table 2- 4 Common kernel functions for SVM

Type of kernel	**Kernel function,** $K(x_i, x_j)$
Linear kernel	$x_i \cdot x_j$
RBF kernel	$\exp\left(-\gamma \left\|\left\|x_i - x_j\right\|\right\|^2\right) \quad (\gamma > 0)$

Polynomial kernel	$(x_i \cdot x_j + 1)^d$

Note: RBF = radial basis function

K is the kernel function.

x is input variables.

γ is a hyperparameter.

d is the degree of the polynomial: d=1 leads to a linear separation, d=2 gives a quadratic kernel, and higher-degree kernels allow a more flexible decision boundary.

SVM is an effective tool in high dimensional spaces, such as document categorization, and generalizes well with high-dimensional data. The ability to apply new kernels allows substantial flexibility for the decision boundaries, leading to greater classification performance.

2.3.3 Decision tree

The decision tree model[83] is constructed in the shape of a tree structure representing the hierarchical organization with multiple branches and levels. Figure 2-3 shows an example of decision tree model. The topmost node, ('age>30') is called the root node. Node without descendants, such as 'minivan' and 'sports car', is called a leaf node, which represents a classification. Other node, such as 'married', is called split node, which has two or more branches.

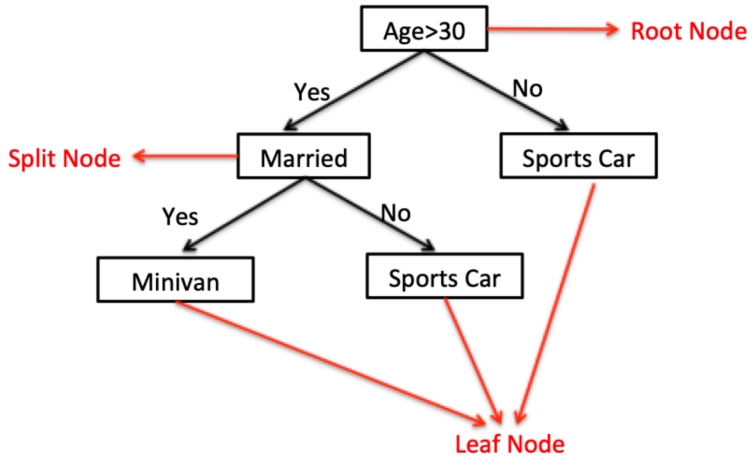


Figure 2- 3 Demonstration of a decision tree

Decision trees can grow very large and complicated. It's easy to create an over-complex classification tree that may fit the training data well, but may do a poor job of classifying new values. This is called overfitting. The process of reducing the size of the tree can reduce model complexity and computation time and is useful for preventing overfitting. Mechanisms such as setting the minimum number of leaf node observations and the maximum number of splits are necessary to avoid this problem. The minimum number of leaf node observations (N) requires that each leaf has at least N observations per leaf node; further splitting of nodes is stopped when the number of observations in the node is lower than N. The maximum number of splits (M) stops further splitting of nodes when the number of split nodes has been reached resulting a tree with M or fewer split nodes. These are called model hyperparameter, a parameter need to be initialized before training a model, which are used to optimize the model performance. For example, kernel function in SVM, value of K in k-nearest neighbor (KNN), depth of tree in decision trees, and number and size of the hidden layers in ANN.

In a dataset, there are many potential variables that can be selected as root node or split node(s). Measures of node impurity are used for selecting the best split. Gini index is a metric to measure how often a randomly chosen element would be incorrectly identified. A variable with lower Gini index is preferred. Gini index for a given node t is calculated as

$$Gini(t) = 1 - \sum_{j} [p(j|t)]^2$$

$p(j|t)$ is the relative frequency of class j at node t.

Another well-known measurement is called entropy which is similar to the Gini index computation. Entropy at a given node t is calculated as

$$Entropy(t) = -\sum_{j} p(j|t) log\ [p(j|t)]$$

2.3.4 Random forest

Unlike a decision tree which builds only one tree, a random forest builds multiple decision trees. It can solve the overfitting problem of a decision tree and run efficiently on large data. Each tree gives a classification (votes for that class). A new object is classified to the class that has the most votes over all the trees. For example, in Figure 2-4, if class A has more votes than class B from the trees, then this new instance is classified as class A. The training set for each tree is created from random resampling of data in original training set with replacement. Therefore, some data could be duplicated and some data could be missing in each training set. This process is called bootstrapping. The size of each training set is the same as the original training set.

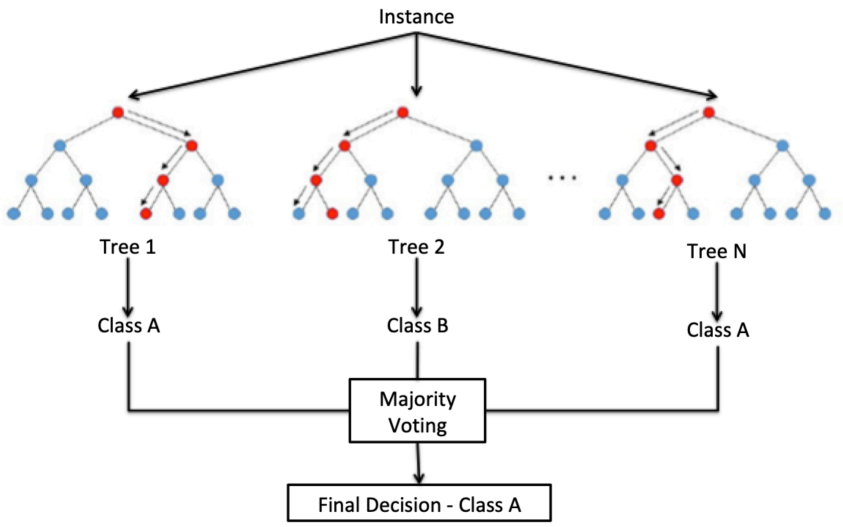


Figure 2- 4 Demonstration of random forest classification

There are some hyperparameters need to be considered when developing random forest. Tree number is the number of trees the model builds before taking the majority voting. In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation. The other hyperparameter is the number of features (NumFeature) randomly selected when splitting a node. If all features are used to build each tree, then this creates a risk of correlation between trees and increases bias in the model. Thus, random forest chooses only a subset of the features at each split to reduce the issue of correlation between trees. As the number of selected features goes up, the strength of the individual trees increases. However, reducing the number of features leads to a lower correlation among the trees and increases the entire model strength.

Out-of-bag (OOB) error is a method of measuring the prediction error of random forests. Suppose the original training data is T and the number of trees are determined as N. N training sets, denoted as $\{T_1, T_2, \ldots, T_N\}$, are created

for N trees $\{K_1, K_2, \ldots, K_N\}$. Each of these training set is called a bootstrap dataset. For each data point D_i in the original training set T, select all the bootstrap datasets that does not contain the data point D_i. This set of bootstrap datasets is called OOB samples. Each data point in the original training set has one set of OOB samples. OOB error is the mean prediction error on each training data point D_i [84]. This can be used to select tree number.

2.3.5 Artificial Neural Network

ANNs are a set of algorithms that were inspired by the human nervous system and work in a similar way to the human brain does. They can be used to extract intricate patterns, discover the relations between the input and output variables, and solve complex problems. ANNs are an assembly of interconnected nodes and weights with three layers: input layer, hidden layer, and output layer (see Figure 2-5). The nodes, which are also known as neurons, are interconnected to process complex information. Each neuron consists of input, weight, and transfer function. A weight is associated with each input and is given to each hidden layer. It represents the strength of its relationship with the output. The hidden layer transforms the input into something that the output layer can use. ANNs process records one at a time and learn by comparing their results with the true labels. As the model develops, the errors from the initial classification (initialized weights) of the first record are used to update the weights and modify the networks algorithm for further iterations. Model accuracy and generalize abilities are improved while training. At last, the output layer sums up each of its input value according to the weights of its links.

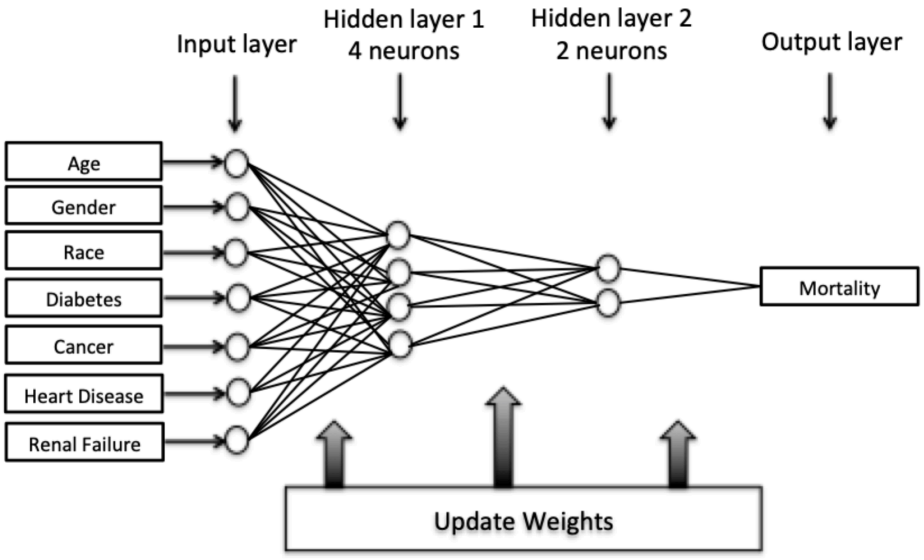


Figure 2- 5 Artificial neural network components

A learning algorithm is applied to train the network between interconnected neurons. Three commonly used learning algorithms include Bayesian regularization (BR) algorithm, Levenberg-Marquard (LM) algorithm, and scaled conjugate gradient (SCG) algorithm. BR updates weight according to BR optimization which minimizes a combination of squared errors and weights, and then determines the correct combination so as to produce a network that generalizes well [85]. LM updates weight according to LM optimization that is designed to reducing error function with a fast and stable convergence [86]. SCG updates the weight and bias values based on conjugate directions without performing a line search at each iteration [87]. LM is recommended for most problems. For some noisy and small problems, BR can take longer but obtain a better solution[88]. SCG is recommended for large problems as it uses gradient calculations which are more memory efficient than the other two algorithms [88].

The advantages and disadvantages of the above machine learning algorithms are shown in Supplementary Table 1.

2.4 Cross validation

Cross-validation (CV) is a common strategy that is used for training and developing model. The entire dataset is divided into k subsets of roughly equal size. Each time, one subset is chosen as the validation set and the rest of subsets are used as training sets (See Figure 2-6 as an example of 5-fold CV). This process is repeated k times, resulting k-fold, and each subset is used exactly once for validation purposes.

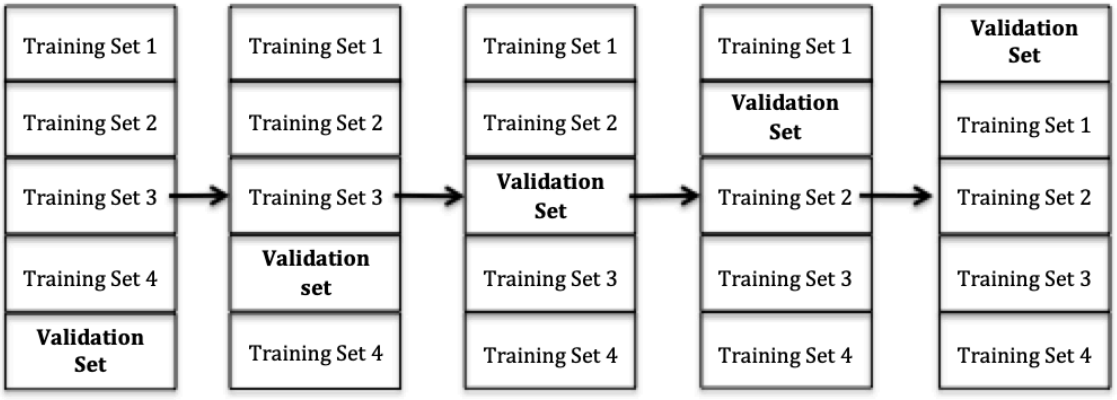


Figure 2- 6 Five-fold cross validation

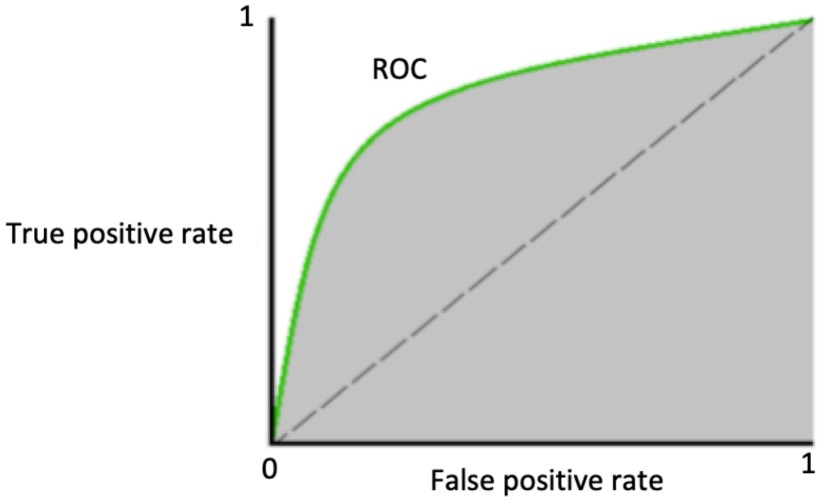
2.5 Model performance metrics

Model performance is tested on validation set during running each fold of cross validation. After k-fold CV, k performance scores will be obtained. The average k performance score is called a CV score. The model with the best CV score, indicating the model with the best predicting ability, is selected as the final optimal model.

The Receiver Operating Characteristic (ROC) curve uses a graphical approach to display classifier performance at all classification threshold (see Figure 2-7). The true positive rate (TPR) and false positive rate (FPR) are plotted on the y-axis and x-axis, respectively, and are defined as follows:

$$TPR = \frac{True\ Positive\ (TP)}{TP + False\ Negative\ (FN)}$$

$$FPR = \frac{False\ Positive\ (FP)}{FP + True\ Negative\ (TN)}$$



ROC = Receiver Operating Characteristic

Figure 2- 7 ROC curve

TPR measures sensitivity, the ability of a model to correctly classify the positive cases. In this project, it would be correctly classifying patients with AEs. It is the proportion of true positives that are correctly classified as positive. High sensitivity means that there are few false negative (FN) results, meaning that fewer positive cases are misclassified. 100% sensitivity indicates that the model correctly classified all positive cases. In this project, a high sensitivity was

required because positive cases are the targeted outcomes which are the patients who had AEs.

Area under the ROC curve (AUC ROC) is another performance measurement used to evaluate the degree of misclassification and is generated as a summary statistic. AUC ROC ranges from 0 to 1 with a higher value indicating a better model ability to distinguish between classes. AUC ROC=0 represents no data point is correctly classified by a model. AUC ROC=1 represents model correctly classifies all data points.

In this project, both sensitivity and AUC ROC were used as evaluation metrics. A combination of a high AUC ROC and high sensitivity were required when selecting the best model. If models have similar AUC ROC and sensitivity, then computational complexity will need to be considered. In this case, the most simple model with low computational complexity should be selected.

Chapter 3 Cohort extraction and preliminary studies

This chapter elaborates the component 1 and component 2 (Figure 1-1) - data cohort extraction, raw data preprocessing, and adverse events risk evaluation in CVD patients who initiated statin agents for secondary prevention.

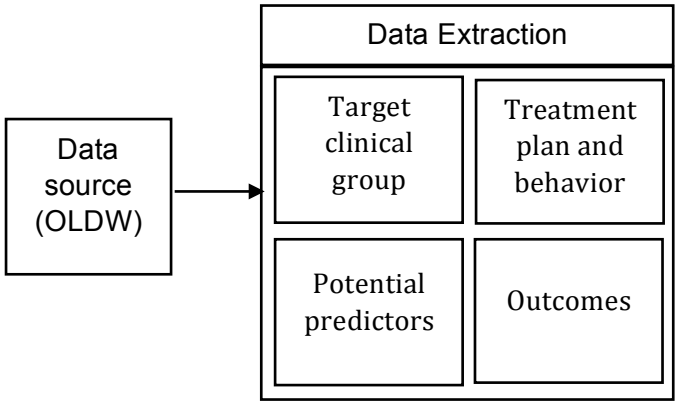
3.1 Introduction

The interest in big-data study is increasing recently. There is a need to extract value from the data to improve patient management and outcomes. To be successful with this work, the key initial challenge was the selection of an appropriate data cohort to support this project. This is especially important for a large-data study because a specific data cohort needs to be extracted from this large database for a specific research focus while also minimizing potential bias. Our original cohort was pulled from the OLDW - a database comprising more than 20 years of data representing more than 200 million de-identified lives, and containing hundreds of variables. Most of the lives in this database did not meet the inclusion criteria of our study and many of the original data elements could not be used directly for this study. In addition, when developing machine learning models, unrelated data elements can cause modeling complexity and negatively influence the ability to learn clear patterns from the data.

This chapter demonstrates how an appropriate data cohort was extracted from the OLDW to optimally support this machine learning and DDIs research. The preliminary results of our patient cohort will also be discussed.

3.2 Methods

A series of tasks were developed to extract an appropriate data cohort from the OLDW (see Figure 3-1). Specific details are discussed as follows.



Note: OLDW = OptumLabs Data Warehouse

Figure 3- 1 Summarization of data-extraction tasks to support the DDI and machine learning research activities.

3.2.1 Target Clinical Group

3.2.1.1. Chronological data inclusion

The OLDW contains multi-year data back to year 1993. Selecting reasonable and manageable years of data is very crucial for study feasibility and generalizability. 2010 was selected as the start year to look for events based on the following reasons. a) Medicare Part D went into effect in 2006, affecting pharmaceutical coverage prices and the utilization of prescription medications with changes in coverage from other insurance plans to Medicare Part D. b) Seven statin agents (atorvastatin, lovastatin, fluvastatin, pravastatin, pitavastatin, simvastatin, and rosuvastatin) were included in this study. Pitavastatin, which is known as Livalo®, is the newest statin in the United States that received FDA

approval in 2009 and was brought to the market in 2010. Therefore, completed prescriptions of all available statins can be obtained starting in 2010.

3.2.1.2. Inclusion and exclusion criteria

Statin claims were selected based on the generic name which includes atorvastatin calcium, fluvastatin sodium, lovastatin, pitavastatin calcium, pravastatin sodium, rosuvastatin calcium, and simvastatin. To support the research purposes, the patient cohort was selected based on the following criteria:

1) Only 40+ years old CVD patients were selected. The prevalence of CVD in patients aged 40 years and above increases substantially from 10% to 40% compared with patients from 20 - 39 years old based on a national survey from 2009 to 2012[89]. In addition, physical condition and adverse event patterns in younger group is different from the older group. Therefore, only patients who were 40-years-old or above and had a stroke, myocardial infarction, and/or coronary revascularization since year 2010 were included. Younger population may be considered added in our future study which include a broader population. Patients with radiation induced coronary artery disease (CAD) or heart transplant related CAD were excluded since the CAD mechanism and treatment strategies are different from other CVD.

2) New CVD patients with no prior CVD history. Patients should not have any CVD diagnose within one year prior to the CVD index date. CVD index date was defined as the first CVD diagnosis date starting from 2010. CVD events were identified by a set of diagnosis and procedure codes (see Supplementary

Table 2) using both inpatient and outpatient codes. Those medical codes were selected by clinical expert review on a published paper [90] and Centers for Medicare and Medicaid Services (CMS) chronic conditions data warehouse (CCW) chronic condition reference list.

3) New statin users who filled an index statin prescription within 30 days after the CVD index date. We required that patient had no statin prescription within one year prior to the statin index date to assure the cohort patients were newly initiated statins. Statin index date was defined as the first statin prescription date after CVD index date. Previous study also required no CVD claims and no statin prescriptions within one year prior to the index date in order to extract new statin and CVD patients[91].

4) All patients were followed up for up to one year after the statin index date to determine the patterns of statin use (continuous, discontinued, or drop out) and observe adverse events. Adverse events typically occur early in the therapeutic course of statin therapy [92,93]. One-year follow-up is likely sufficient to capture most of the clinical events and made it easier to conduct comparisons in subsequent work.

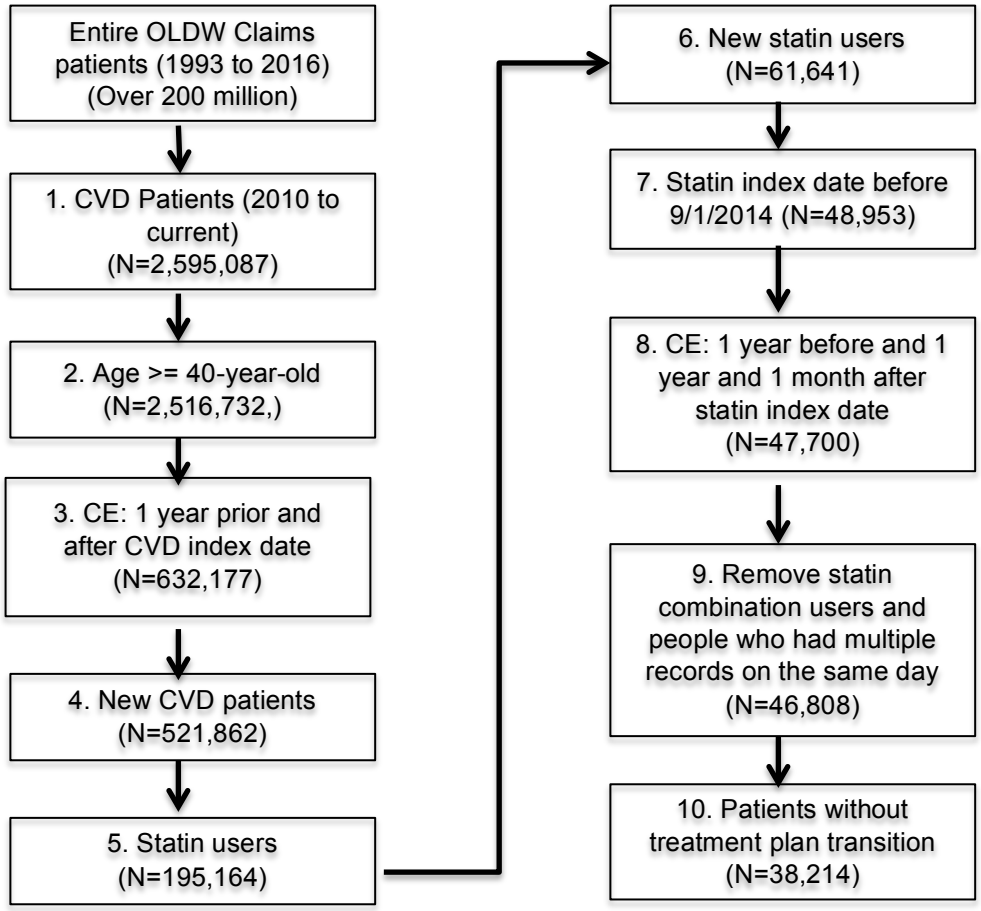
5) Statin index date had to be prior to 9/1/2014. Cohort extraction work was processed in early 2016. Medical claims in OLDW less than 6 months old can still be in the process of adjudication and have yet to be finalized. Therefore, I only used medical claims before September 2015 in order to get complete medical claims. Using this criteria, the statin index dates should be between 1/1/2010 and 8/31/2014.

6) Patients were required to have continuous medical and pharmacy enrollment one year prior to the CVD index date and one year and one month after statin index date. The reason that an extra one month of continuous enrollment was needed after one-year follow-up is that I used 30-day window to differentiate pattern of statin use (details are described in 'treatment plan and behavior' section). Therefore, one extra month of statin claims was needed to categorize some patients.

7) To reduce the bias due to treatment plan transition, only patients who maintained the same treatment plan were included (consistent statin agent and dosage; see Table 3-3). In other words, this patient group only had one treatment plan before statins were discontinued, an AE occurred, or the end of the one-year follow-up period was attained.

8) Patients who used combination drugs including amlodipine/atorvastatin, ezetimibe/atorvastatin, ezetimibe/simvastatin, niacin/simvastatin, sitagliptin/simvastatin, lovastatin/niacin, and aspirin/pravastatin were excluded. Specifically, combination drugs include two or more drugs in the same tablet, of which one is a statin agent. To keep the cohort patients more generalizable, those patients were excluded from this study because the combination medications may have different pharmacodynamics and patterns of use for patients receiving therapy. Because the statins are frequently given at different times than other medications, such as with dinner or before bed, having them combined with another medication simultaneously may be problematic. For instance, if a participant experiences a side effect due to the non-statin

component of a combination drug, it would change their medication experience, affect patterns of AEs, and potentially make predictive modeling more difficult and less generalizable for statin medications. Flowchart of study cohort selection and development is shown in Figure 3-2.



Note: OLDW = OptumLabs Data Warehouse

CVD = cardiovascular disease

CE = continuous medical and pharmacy enrollment

Figure 3- 2 Flowchart for selection of the study population from OLDW

3.2.2. Potential Predictors in predictive model

The predictive model is intended to predict the risk of adverse events. Predictors (variables) that can be obtained from the OLDW were included. Specifically, these variables include demographic variables, type of insurance (commercial, Medicare Advantage), medical cost (both patient and health plan paid amounts), medication costs (both patient and health plan paid amounts), administrative claim comorbidities, provider specialty, and other variables that can be collected in the OLDW (See details in Supplementary Table 3).

3.2.3. Treatment Plan and Behavior

Previous studies have shown that the risk of AEs is associated with specific statin agents[34,94,95] and a statin dose effect[21,96,97]. Statin were categorized into three intensity groups (low, moderate, and high intensity), based on the 2013 American College of Cardiology/American Heart Association (ACC/AHA) guideline on the treatment of blood cholesterol[6] (see Table 3-1). In addition, patients' statin tolerance behaviors were classified into three groups, continuous, discontinued, and dropout, to facilitate observation of the extracted data. People who continuously used statin treatment for at least one year without any gap longer than 30 days were categorized as continuous group. Patients who had only one statin prescription during the study period were categorized as dropout group. Patients in the discontinued group are those who had a gap more than 30 days or did not continuously use statin treatment for one full year. Gap was calculated as follows:

Gap days=2nd statin fill date - (1st statin fill date + 1st fill of days of supply)

Table 3- 1 Statin intensity groups

Statin	Intensity		
	Low	Moderate	High
Atorvastatin	--	10mg, 20mg	40mg, 80mg
Simvastatin	5mg, 10mg	20mg, 40mg	80mg
Pravastatin	10mg, 20mg	40mg, 80mg	--
Rosuvastatin	--	5mg, 10mg	20mg, 40mg
Lovastatin	10mg, 20mg	40mg	60mg
Pitavastatin	1mg	2mg, 4mg	--
Fluvastatin	20mg, 40mg	80mg	--

3.2.4. Outcome measurements

The statin related AEs are the main outcomes in the drug-drug interaction study and the predicted outcome in the treatment plan predictive model. They were identified by International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). In this project, only major clinically important AEs were considered including rhabdomyolysis**,** myopathies, renal events, liver events, and medication poisoning events associated with statins (See Table 3-2). Some minor AEs, such as muscle pain, were not considered, since they may not have consistent ICD-9 codes and likely only exist in clinical notes in many cases creating risk of misclassification.

Incidence rates (IRs) of statin adverse events were used to evaluate the adverse event results. It was calculated as the number of incident cases divided

by the total person-time and was reported as the number of cases per 100 person-years. Person-time was calculated as the time when patients were at risk of adverse events during the study period. Person-time was censored at the point that patient experienced an adverse event, stopped using statins, or reached the end of the observation time of the study, whichever happened first.

Table 3- 2 Selected ICD-9 codes of adverse events

Adverse events	**ICD-9 Codes**
Rhabdomyolysis	728.88 [98]
Myopathies	359.4, 359.8, 359.9, 728.8, 728.89, 728.9, 729.1, 791.3 [24,99]
Acute kidney injury	584.XX [100,101]
Liver events	570, 573.X [102,103]
Poisoning events	972.2, E942.2, E980.4 [24]

3.3 Results

We had more than 2 million CVD patients who were 40-year-old or above in the OLDW (Figure 3-2). After imposing one-year continuous medical and pharmacy enrollment before and after the CVD index date, there were 632,177 patients left. Then patients who had any CVD diagnosis or statin prescription within the 1-year drug and disease free period were removed and 61,641 new CVD and statin users were obtained. Then 12,688 patients whose statin index dates were after August 31, 2014 were excluded. After imposing continuous enrollment of one year before and one year and one month after statin index date, 47,700 patients included. To keep the cohort simple and pure, 892 people were

excluded because of using statin combination therapy or having multiple statin prescriptions on the same day. Among the 46,808 patients, 8,594 patients either switched to another statin agent or modified the statin dose. In all, 38,214 patients were included in the final cohort population who adhered to the same statin treatment plan within the observed time period. 16,333 (42.7%) patients had Medicare Advantage plans and the rest had commercial plans. Table 3-3 shows the patient numbers for each number of statin treatment plans and number of patients who maintained the same statin agent but decreased or increased the statin strength.

Table 3- 3 Summary of number of statin treatment plans change during the study period

# Statin treatment plan changes	Count, n (%)	Dose modification, n	Switch in statin agent, n
0	38,214 (81.6)	NA	NA
1	7,417 (15.8)	4,342	3,075
2	1,022 (2.2)	273	749
3	135 (0.3)	<11	>124
4+	20 (0)	0	20

Table 3- 4 Demographic characteristics for patients in each statin group

Patient Group by Statins	Patients Number, N (%) (Total N=38,214)	Age (Mean ± SD)	Male (%)	Intensity Groups		
				Low, N	Moderate, N	High, N
Atorvastatin	15,440 (40.4)	63.5±11.6	64.2	NA	6,976	8,464
Simvastatin	>13,801 (36.1)	64.9±11.7	58.7	1,750	11,449	>602
Pravastatin	5,210 (13.6)	65.5±11.6	55.2	2,314	2,896	NA
Rosuvastatin	3,223 (8.4)	61.1±10.8	63.6	NA	1,980	1,243
Lovastatin	433 (1.1)	66.3±11.7	53.8	311	122	0
Pitavastatin	92 (0.2)	61.7±10.0	57.6	17	75	NA
Fluvastatin	<15 (<0.04)	57.9±9.2	*	<11	<11	NA

Note: * Due to the small size policy, percentage is masked.

The mean age of the cohort population was 64.1 years (sd=11.7 years). The mean age for each statin agent group ranged from 57.9 years in fluvastatin group to 66.3 years in lovastatin group (Table 3-4). Atorvastatin was the most popular prescribed statin agent (40.4%), followed by simvastatin (36.1%). The gender distribution and the number of patients in each intensity group are also listed in Table 3-4.

Table 3- 5 The incidence rate for each adverse event by different statin agent groups

Statins	Incidence Rates (per 100 person-years)					
	Any AEs	Rhabdomyolysis	Myopathy	Renal	Liver	Poisoning
Atorvastatin (n=15,540)	29.3 (28.1-30.5)	0.15 (0.07-0.23)	19.0 (18.1-19.9)	5.9 (5.4-6.4)	5.1 (4.6-5.5)	0.16 (0.08-0.25)
Simvastatin (n>13,801)	30.0 (28.7-31.3)	0.26 (0.15-0.37)	20.4 (19.3-21.4)	6.0 (5.5-6.6)	4.4 (3.9-4.8)	0.05 (0-0.10)
Pravastatin (n=5,210)	32.2 (30.0-34.5)	0.14 (0.00-0.28)	21.7 (19.9-23.6)	6.4 (5.5-7.4)	4.4 (3.6-5.2)	0.14 (0-0.28)
Rosuvastatin (n=3,223)	25.3 (22.7-27.9)	0.25 (0.01-0.50)	17.5 (15.4-19.6)	3.8 (2.8-4.7)	4.0 (3.0-5.0)	0.06 (-0.06-0.19)
Lovastatin (n=433)	23.9 (17.1-30.7)	0.46 (-0.44-1.36)	18.2 (12.4-24.1)	4.7 (1.8-7.5)	3.2 (0.8-5.6)	0
Pitavastatin (n=92)	25.6 (7.9-43.3)	0	25.6 (7.9-43.3)	0	3.0 (-2.9-8.9)	0
Fluvastatin (n<15)	0	0	0	0	0	0

Note: AE = adverse event

Table 3-5 summarizes the incidence rates for specific adverse event by different statin agent group. Each individual adverse event as well as the combination of all AEs were investigated. The IRs of combined adverse events ranged from 25.3 cases (rosuvastatin) to 32.2 cases (pravastatin) per 100 person-years. For rhabdomyolysis, lovastatin users had the highest IR (0.46 cases per 100 person-years). Pravastatin and atorvastatin users had similar IRs (0.14 and 0.15 cases per 100 person-years, respectively). Rosuvastatin and simvastatin also had very similar IRs (0.25 and 0.26 cases per 100 person-years, respectively). For myopathy, rosuvastatin group had the lowest IR (17.5 cases per 100 person-years), while pitavastatin had the highest IR (25.6 cases per 100 person-years). For renal events, the incidence rates ranged from 3.8 cases (rosuvastatin) to 6.4 cases (pravastatin) per 100 person-years. For liver events, pitavastatin had the lowest IR (3.0 cases per 100 person-years) and atorvastatin had the highest IR (5.1 cases per 100 person-years). IRs of poisoning events were between 0.05 cases (simavastatin) to 0.16 cases (atorvastatin) per 100 person-years.

The entire cohort was then divided into three groups - continuous, discontinued, and dropout. Among the cohort patients, 39.4% maintained the same statin regimen for at least one year (continuous), 13.3% had only one statin prescription (dropout), and the remainder of patients discontinued statin treatment within one year (discontinued) (see Table 3-6). The duration of exposure to statins was also calculated during the study period. The average continuous statin exposure time for the continuous, discontinued and dropout

group were 11.8 months, 4.2 months, and 1.2 months, respectively. Comparisons were further conducted after stratifying by individual statin (Table 3-7). Pitavastatin has the highest discontinuation and dropout rate. Rosuvastatin and lovastatin have moderate discontinuation and dropout rate. Atorvastatin, simvastatin and pravastatin have the lowest discontinuation and dropout rate.

Table 3- 6 Demographic characteristics based on tolerance group

Tolerance group	Total patients (%)	Age (mean ± SD)	Male (%)	Total continuous statin exposure time (years)
Continuous	39.4	65.0±11.4	64.3	14793.7
Discontinued	47.3	63.4±11.7	60.2	6303.7
Dropout	13.3	64.3±12.2	52.9	514.9

Table 3- 7 Demographic characteristics for each tolerance group stratified by individual statin agent

Statin agents	Tolerance group	Total patients (%)	Age (mean ± SD)	Male (%)	Total continuous statin exposure time (years)
Atorvastatin	Continuous	42.7	63.9±11.3	67.6	6,484.2
	Discontinued	46.0	62.9±11.7	63.0	2,569.0
	Dropout	11.3	64.6±12.5	56.6	173.3
Simvastatin	Continuous	38.9	66.2±11.4	62.0	5,276.0
	Discontinued	46.8	64.0±11.8	58.3	2,273.9
	Dropout	14.3	64.3±12.2	50.9	198.4

Pravastatin	Continuous	36.9	67.4±11.3	56.9	1,890.2
	Discontinued	47.0	64.4±11.6	56.3	817.0
	Dropout	16.0	64.8±11.8	48.1	88.4
Rosuvastatin	Continuous	31.4	60.8±10.2	69.5	990.5
	Discontinued	54.5	61.0±10.8	61.9	550.4
	Dropout	14.1	62.5±12.1	57.0	44.4
Lovastatin	Continuous	31.2	68.8±10.7	53.3	133.1
	Discontinued	53.6	65.5±12.2	54.3	77.0
	Dropout	15.2	64.1±11.0	53.0	8.0
Pitavastatin	Continuous	18.5	60.9±9.0	*	16.7
	Discontinued	57.6	61.5±9.8	62.3	14.5
	Dropout	23.9	62.9±11.3	*	2.0
Fluvastatin	Continuous	*	52.7±11.5	*	*
	Discontinued	*	61.6±10.0	*	1.9
	Dropout	*	57.0±3.6	*	0.4

Note: * Due to the small size policy, percentages are masked.

Table 3- 8 Incidence rates for each adverse event group stratified by statin tolerance group

Type of AEs (N)	Tolerance Groups	Number of AE patients (N)	Total person-years	IRs (per 100 person-years)
Any AEs (N=5,640)	Continuous	2,695	12,950	20.8 (20.0-21.6)
	Discontinued	2,489	5,659	44.0 (42.3-45.7)
	Dropout	456	483	94.5 (85.8-103.1)

Rhabdomyolysis (N=43)	Continuous	>10	*	0.09 (0.04-0.14)
	Discontinued	22	6,300	0.35 (0.20-0.50)
	Dropout	<11	*	1.6 (0.5-2.6)
Myopathy (N=3,920)	Continuous	1,793	13,535	13.2 (12.6-13.9)
	Discontinued	1,784	5,839	30.6 (29.1-32.0)
	Dropout	343	491	69.9 (62.5-77.3)
Renal events (N=1,233)	Continuous	597	14,411	4.1 (3.8-4.5)
	Discontinued	563	6,174	9.1 (8.4-9.9)
	Dropout	73	510	14.3 (11.0-17.6)
Liver events (N=980)	Continuous	526	14,481	3.6 (3.3-3.9)
	Discontinued	395	6,213	6.4 (5.7-7.0)
	Dropout	59	511	11.6 (8.6-14.5)
Poisoning events (N=24)	Continuous	*	*	0.11 (0.06-0.16)
	Discontinued	*	*	0.06 (0.00-0.13)
	Dropout	*	*	0.78 (0.02-1.54)

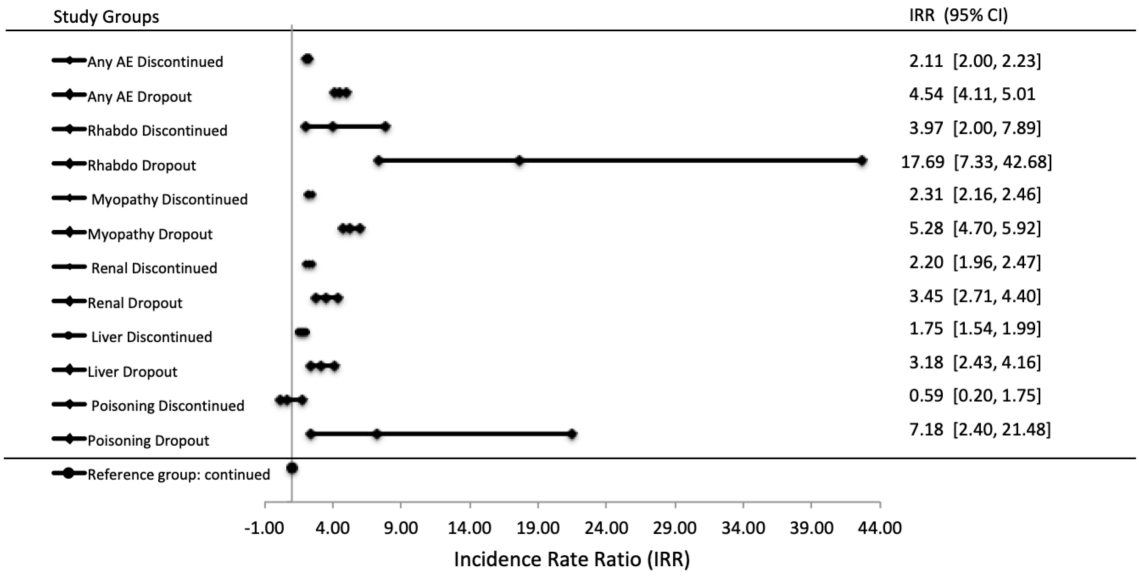
Note: AEs = adverse events.

IRs = incidence rates.

*: Due to the small size policy, numbers are masked. Some total person-years are also masked because the number of patients can be back-calculated using IR and total person years.

Table 3-7 summarizes the observation in terms of number of AE patients, total person-years, and the IRs. IR of continuous group was then used as the

reference group to calculate and compare the incidence rate ratio (IRR) (see Figure 3-3).



Note: IRR = incidence rate ratio; CI = confidence interval

Figure 3- 3 Incidence rate ratio for each adverse event group stratified by statin tolerance group as compared to continuous group

For overall adverse events, the average person-years each patient contributed to the continuous, discontinued, and dropout groups were 0.98 person-years, 0.35 person-years, and 0.1 person-years, respectively. The IRs of continuous, discontinued, and dropout group were 20.8, 44.0, and 94.5 cases per 100 person-years (Table 3-7). The IRs of discontinued and dropout groups were significantly 2.11 times (95% CI, 2.00-2.23) and 4.54 times (95% CI, 4.11-5.01) greater than that of the continuous group (Figure 3-3).

For rhabdomyolysis, myopathy, renal events, and liver events, the continuous group had the lowest IRs (0.09 cases, 13.2 cases, 4.1 cases, and 3.6 cases per 100 person-years, respectively) and dropout group had the highest IRs

(1.6 cases, 69.9 cases, 14.3 cases, and 11.6 cases per 100 person-years, respectively). The IRs for both discontinued and dropout groups were significantly higher than that of in the continuous group.

For poisoning events, the IRs ranged from 0.06 cases per 100 person-years in the discontinued statin users group to 0.78 cases per 100 person-years in the dropout users group. The IR of the discontinued group was 0.59 times of the continuous group, but was not statistically different (95% CI [0.20 – 1.75]). The IR of the dropout group was significantly 7.18 times higher than that of in the continuous group (95% CI [2.4 – 21.5]).

3.4 Discussion

To improve the chance of success of the entire project, I extracted an appropriate patient cohort to investigate drug-drug interaction to facilitate development of prediction models.

Preliminary tests were conducted to get a general idea about the characteristics of the cohort patients. 38,214 patients maintained the same statin treatment plan within one year and were selected as the final cohort. Among those patients, 40.4% of patients used atorvastatin, followed by simvastatin which is the second most prescribed statin representing 36.1% of the cohort patients. Nearly 40% of patients continuously used a statin for at least one year. The rest of the patients either discontinued the statin after a period of time or stopped the statin after their first statin prescription. This is consistent with previous studies[10,27]. The average time of statin therapy for patients in the

continuous, discontinued, and dropout groups were 11.8 months, 4.2 months, and 1.2 months, respectively.

Individual AEs as well as the combination of all types of AEs were also investigated. Patients in discontinued and dropout groups generally had higher IRs than continuous group. AEs in these groups may contribute to patient's decision to discontinue statins. When AEs were compared by each statin agent group, pravastatin was found to have the highest IRs for combination AEs and renal events. Pitavastatin users had the highest IR of myopathy. Atorvastatin users had the highest IRs of liver and poisoning events. Lovastatin users had the highest IR of rhabdomyolysis. Myopathy was the most common adverse event. Renal and liver events had much lower IRs. Rhabdomyolysis and the poisoning events were very rare.

Limitations in this section of the study included: 1) Only patients without statin treatment transitions were selected to be in the patient cohort, which reduces generalizability. However, this step was necessary to reduce possible bias in AE prediction due to changes in treatment plan. 2) Some AEs, such as mild muscle pain cannot be fully captured since patients may not be seen in the clinic, emergency department for mild pain, or it may only be noted in physician notes . Thus, only AEs with ICD-9 codes were selected. 3) The defined follow-up period was one year. If the follow-up period was longer, more AEs may have been detected, and vice versa. AE patterns may be different with different lengths of follow-up. 4) Affordable Care Act (ACA), which was enacted in 2010, changed healthcare in many aspects, such as increasing insurance coverage,

changing insurance standards, and affecting insurance premiums and healthcare cost. This would suggest that 2010 might be an unstable year as a lot of patient transitions occurred that year.

Chapter 4 Simvastatin drug-drug interaction

This chapter includes component 3 (Figure 1-1) - evaluation on the effect of the combination therapy of simvastatin and several pre-defined high risk interacting drugs (CYP3A4 and/or OATP inhibitors) in CVD patients using simvastatin medications for secondary prevention.

4.1 Introduction

Drug-drug interactions are a common cause of AEs which are responsible for up to 2.8% of hospital admissions[1]. Many CVD patients may need statins as well as other therapies, especially in those who have multiple comorbidities, and those at high CVD risk who cannot achieve optimal therapeutic benefits from statin monotherapy.

In this chapter, I focused on simvastatin drug-drug interaction as simvastatin is one of the earliest statins that was approved by FDA and has a long history of medical application. It is one of the most commonly prescribed statin agents. According to our data, simvastatin (36.1%) and atorvastatin (40.4%) are the two most prescribed statin agents. They are both metabolized by the enzyme cytochrome P450 3A4 (CYP3A4). However, simvastatin undergoes more pre-systemic metabolism than atorvastatin[104] which results in lower bioavailability for simvastatin (<=5%) compared with atorvastatin (12%). Drugs with high intestinal and liver extraction are often involved in significant DDIs when concomitant use with enzyme inhibitors or inducers. Therefore, simvastatin is more susceptible to medicine interactions. Study showed that simvastatin's blood levels may be increased five-fold or higher by CYP3A4 inhibitors[105]. In addition,

FDA restricted the use of the highest approved dose of simvastatin (80mg) because of its increased risk of muscle related events[106]. FDA recommended that “simvastatin 80 mg should be used only in patients who have been taking this dose for 12 months or more without evidence of muscle injury (myopathy)”. Given the innate characteristics of simvastatin, its long history of use and large number of individuals using the medication, it is an important medication on which to assess the risk of adverse events associated with DDI exposures.

The purpose of this chapter was to evaluate the effect of combination therapy of simvastatin and several pre-defined high risk interacting drugs, which are CYP3A4 and/or OATP inhibitors, on CVD patients in a large administrative claims dataset. Many case reports[51,70,107,108] of DDI associated AEs involved concomitant use of simvastatin and CYP3A4 or OATP1B1 inhibitors. A few population-based studies[50,109–111] investigated statin DDIs. However, they either focused on composite statin use or composite interacting drugs which cannot be used to determine the interactions induced by a specific statin and specific interacting drug. Our study extends these findings by focusing on one specific statin (simvastatin) and several predefined high risk interacting drugs. To better understand the interactions between simvastatin and specific interacting drugs, three drugs which were known metabolic inhibitors were selected for subgroup analysis due to their relatively large sample size with adequate power for statistical analysis. In addition, to the best of our knowledge, no published population-based study specifically differentiated among different time periods: pre-DDI, DDI exposure, and post-DDI time periods. Comparisons between the

DDI-exposed group and non-DDI-exposed groups as well as the comparisons within subjects among those with DDI-exposure including pre-DDI, DDI exposure, and post-DDI time frames were performed.

4.2 Methods

4.2.1. Study population

The patients for DDI study were selected from the cohort patients who were prescribed simvastatin. They were divided into two groups according to whether they were exposed to the predefined high risk interacting drugs: DDI group (exposed to an interacting drug) and non-DDI group (not exposed to an interacting drug). DDI group patients were identified when concomitant administration of at least one interacting drug during the simvastatin exposure period. The concomitant medication was defined as occurring when the prescriptions of simvastatin and the interacting drug had an overlapping exposure period.

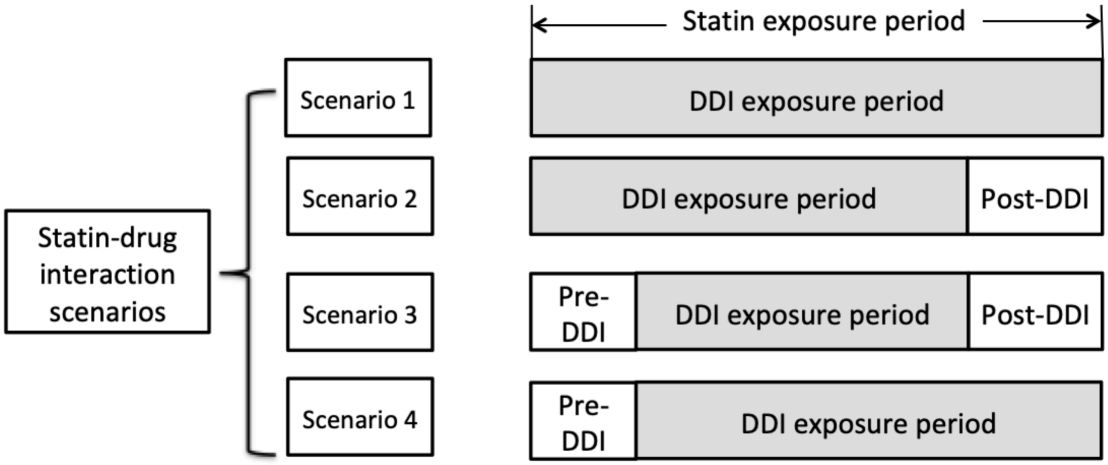
4.2.2. Interacting drugs

The interacting drugs were selected by clinical expert review after carefully screening available DDI resources including drugs.com, Lexicomp, Epocrates, and published literature[49]. Drugs that had the highest DDI risk across each of the DDI resources were selected, including clarithromycin, telithromycin, erythromycin, nefazodone, itraconazole, ketoconazole, posaconazole, boceprevir, danazol, cobicistat, gemfibrozil, and cyclosporine. These medications are CYP3A4 and/or OATP inhibitors and had been previously identified as increasing risk of AEs with simvastatin.

To investigate the safety of specific interacting drugs, a subgroup analysis was conducted among three interacting drugs - gemfibrozil, clarithromycin, and erythromycin, as they had a relatively large number of patients with adequate power for statistical analysis.

4.2.3. Study time frames definition

For patients in the DDI group, the study period was divided into three time frames: 1) pre-DDI which was prior to initiation of DDI agent, 2) DDI exposed which was while the patient took simvastatin and an interacting medication without any gap longer than 30 days, and 3) post-DDI which was after the DDI exposure ended. AEs were detected during each of the above time periods.



Note: DDI = drug-drug interaction

Figure 4- 1 Scenarios of statin-drug interaction and illustration of pre-DDI and post-DDI period for each scenario

4.2.4. Outcome measures

Major clinically important AEs that were discussed in Chapter 3, including rhabdomyolysis, myopathies, renal adverse events, hepatic adverse events, and

statin poisoning events, were included.

Incidence rates were used to evaluate the risk of AEs in different groups by accounting for the potential difference of length of drug exposure. IRs were reported as the number of cases per 10 person-years of exposure.

The number of physician claims per month was used as another study outcome to assess if patients with DDI exposure had increased clinical follow-up to manage the potential DDI risk compared with non-DDI group. This study hypothesized that subjects exposed to the interacting drugs would have more intense medical follow-up. Physician claims on the same day were counted as one claim. Outpatient claims, emergency room (ER) visits, and office visits were included in the physician claim totals.

In addition, Charlson comorbidity index score was compared between groups. The Charlson comorbidity index is the most widely used for predicting one-year mortality based on comorbidity data [112]. It has also been used as a predictor for adverse events[113–120]. The index score was used instead of single comorbidities because it is a summary comorbidity measure reflecting risk of death from many comorbid diseases. A score of zero indicates that no comorbidities were found. The higher the score, the more likely that death will occur.

4.2.5. Statistical analysis

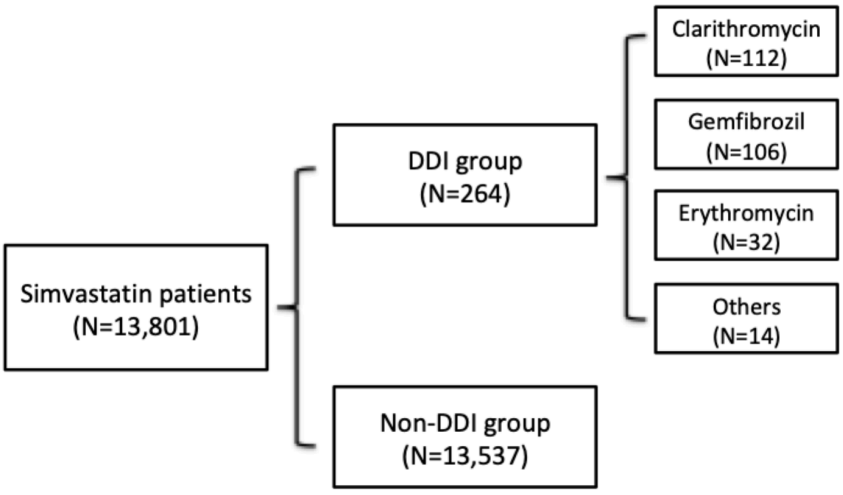
R software was used for analysis. Two-tailed t-students tests were performed for continuous variables. 95% confidence intervals (CI) were

calculated for incidence rate ratio. Analyses were carried out with RStudio version 0.98.1103. A p-value < 0.05 was considered statistically significant.

4.3 Results

13,801 simvastatin CVD patients, who did not have any interacting drug or had only one interacting drug, were included in this study. Patients who took two or more different interacting drugs during simvastatin exposure time period were excluded because they had a complex pre-DDI and post-DDI period and a very small sample size (N<11).

To protect patient privacy and confidentiality, small numbers or small percentages which represent small numbers (N<11) cannot be reported according to the OptumLabs cell size suppression policy.



Note: DDI = drug-drug interaction

Figure 4- 2 Patient number for non-DDI group, DDI group, and sub-DDI groups

A total of 264 patients were identified in the DDI group including 112 clarithromycin patients, 106 gemfibrozil patients, 32 erythromycin patients, and a total of 14 patients who took ketoconazole, nefazodone, cyclosporine, or

itraconazole. No patient used telithromycin, posaconazole, boceprevir, danazol, or cobicistat. Figure 4-2 shows the patient number in each group and sub-group.

4.3.1 Baseline comparison between DDI group and non-DDI group

A comparison of baseline patient characteristics between simvastatin DDI group and non-DDI group is shown in . From the table, we can see DDI group had a significantly longer statin exposure time and higher Charlson score than non-DDI group. Table 4-2 shows IRs comparison between DDI and non-DDI group. DDI group had a 1.29 times higher incidence risk of AEs compared with the non-DDI group after adjusting for statin exposure time. (See Supplementary Figure 1 for key dates, time period and outcomes illustration.)

Table 4- 1 Patient characteristics comparison between DDI and non-DDI groups

Patient characteristics	DDI group (N=264)	Non-DDI group (N=13,537)	p-value
Statin exposure time (days)	249.1	204.0	<0.0001*
Age (years)	63.9	64.9	0.1555
Male (%)	162 (61.4%)	7,934 (58.6%)	0.3646
Charlson index score	1.55	1.10	0.0002*
Intensity (%)			
Low	30 (11.4%)	1,719 (12.7%)	0.5003
Medium	220 (83.3%)	11,226 (82.9%)	0.8616
High	14 (5.3%)	592 (4.4%)	0.505

Note: DDI = drug-drug interaction

* indicates results are significantly different.

Table 4- 2 Incidence rate ratio comparison between DDI and Non-DDI groups

	AE counts [c]	IR [a]	IRR [b] [95% CI]
Non-DDI group (N=13,537)	1,957	2.84	---
DDI group (N=264)	13	3.66	1.29 [0.75, 2.22]

Note: DDI = drug-drug interaction

a. IR = incidence rate.

b. IRR = incidence rate ratio.

c. In DDI group, AEs were counted when they occurred during DDI exposure period. In Non-DDI group, AEs were counted when they occurred during the statin exposure study period.

4.3.2 Physician claims comparisons

Table 4-3 shows the results of the numbers of physician claims per month. Patients who did not have any physician claims were counted as contributing zero visits. Among the patient cohort, 642 patients (4.7%) resulted zero visits. Since only a small percentage of the cohort had zero visits, they were included in the following analysis.

The results show that the DDI group had a significantly higher number of physician claims than the non-DDI group (2.53 claims/month vs. 2.17 claims/month with p=0.0060). Similar analyses were done among the DDI group patients. 38 of 264 patients exposed an interacting drug throughout the whole simvastatin period (scenario 1 in Figure 4-1) resulting in zero days of DDI unexposed time for these patients. The number of physician claims per month

during the DDI exposed and unexposed time periods were 3.95 and 2.38 with a statistically significant p-value less than 0.0001. A significant result was also found when comparing the two DDI unexposed time periods. Pre-DDI period had a higher number of physician claims (3.22 claims per month) compared with the post-DDI period (1.91 claims per month).

Table 4- 3 Physician claims comparisons between different groups

Patients	Groups	Claim numbers per month (N)	p-value
All patients	DDI group	2.53 (N=264)	0.0060*
	Non-DDI group	2.17 (N=13,537)	
DDI patients	DDI exposed period	3.95 (N=264)	
	DDI unexposed period**	2.38 (N=226)	<0.0001*
DDI patients	Pre-DDI	3.22 (N=164)	<0.0001*
	Post-DDI	1.91 (N=196)	

Note: DDI = drug-drug interaction

* Results are significantly different.

** DDI unexposed period includes both pre-DDI and post DDI periods.

4.3.3 Baseline comorbidities comparisons

The above IRR and physician claim comparison did not adjust for comorbidities, thus we compared individual Charlson comorbidity between DDI and non-DDI group to get a better sense whether the differences may be able to explained by underlying comorbidity or are more likely to be explained by DDI. Table 4-4 shows DDI group has significantly higher percentage of patients with

pulmonary disease, rheumatologic disease, and diabetes (mild to moderate). However, liver and renal disease, which could potentially affect the AEs during medication exposure, does not show statistically significance.

Table 4- 4 Incidence rate ratio comparison between DDI and Non-DDI groups

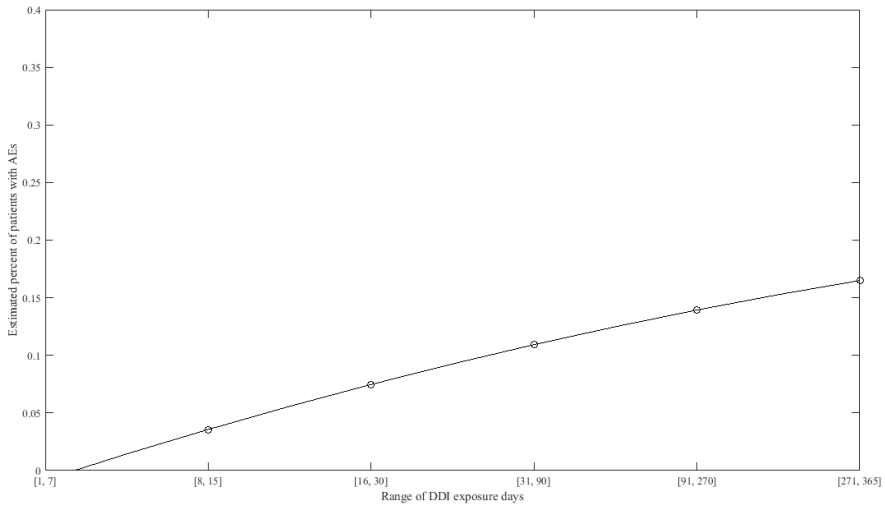
Charlson comorbidities	DDI group (N=264)	Non-DDI group (N=13,537)	p-value
Myocardial infarction	<4.2%	3.7%	NS[1]
Renal disease	4.9%	4.9%	0.9974
Moderate or severe liver disease	<4.2%	0.14%	NS[1]
Mild liver disease	4.9%	3.1%	0.1657
Congestive heart failure	5.7%	6.4%	0.6240
Peripheral vascular disease	11.7%	8.1%	0.0687
Cerebrovascular disease	8.3%	7.8%	0.7441
Dementia	<4.2%	0.92%	NS[1]
Chronic pulmonary disease	23.5%	16.1%	0.0052*
Rheumatologic disease	9.1%	2.8%	0.0005*
Peptic ulcer disease	<4.2%	0.85%	NS[1]
Diabetes (mild to moderate)	31.4%	18.9%	<0.0001*
Diabetes with chronic complications	6.8%	4.4%	0.1296
Paraplegia or hemiplegia	<4.2%	0.48%	NS[1]
Any malignancy, including	9.1%	7.8%	0.4839

lymphoma and leukemia			
Metastatic solid tumor	<4.2%	0.73%	NS[1]
AIDS	<4.2%	0.15%	NS[1]

Note: 1: NS: not significant. Small numbers or small percentages which represent small numbers (N<11) cannot be reported according to the OptumLabs cell size suppression policy. The corresponding P-values are replaced by NS if p-value is >0.05.

4.3.4 Relationship between DDI exposure time and adverse event rates

The estimated percentages of patients who had any type of AEs during DDI exposure period are shown in Figure 4-3. Individual patients contribute only 1 AE if AE(s) occurred. A temporal association of percentage of AE patients with DDI exposure time was assessed. DDI group patients were divided into six groups based on the length of DDI exposure time (1-7 days, 8-15 days, 16-30 days, 31-90 days, 91-270 days, and 271-365 days) to provide a general time frame for the occurrence of AEs. This figure indicates the number of people who have AE increases as the DDI exposure time increases.



Note: AEs = adverse events

DDI = drug-drug interaction

Figure 4- 3 Estimated percentage of AE patients with DDI exposure time

4.3.4 Subgroup analysis in three specific interacting drugs

Table 4-5 shows the results of the three specific interacting drugs: gemfibrozil, clarithromycin, and erythromycin. IRs were measured in pre-DDI, DDI exposure, and post-DDI time periods. Person time was calculated as the total time from the start date to the date of index AE occurrence or to the end of time period, whichever occurred first. IRRs were calculated using IR of non-DDI group (2.84 per 10 person-year) as the reference group. Both clarithromycin and erythromycin have higher IRs in all three time frames compared with non-DDI group. Gemfibrozil DDI group had significantly higher IR in the pre-DDI and DDI exposure period, but lower IRs in the post-DDI period (not significant) compared with the non-DDI group. The risks of AEs for clarithromycin and erythromycin were higher than gemfibrozil during DDI exposure and post-DDI time period.

However, gemfibrozil had the highest pre-DDI IR compared with the other two inhibitors. The highest IRs of clarithromycin and erythromycin occurred in the DDI exposed period. However, the highest IR of gemfibrozil occurred in the pre-DDI time period.

Table 4- 5 Risk estimate in concomitant use of simvastatin and interacting drugs

Interacting drugs (N)	Time frames	IRs	IRRs** [95% CI]
Gemfibrozil (N=106)	Pre DDI	8.84	3.11 [1.17, 8.30]*
	DDI exposure	3.11	1.10 [0.57, 2.11]
	Post DDI	1.51	0.53 [0.23, 1.63]
Clarithromycin (N=112)	Pre DDI	3.26	1.15 [0.65, 2.02]
	DDI exposure	6.09	2.14 [0.54, 8.58]
	Post DDI	3.64	1.28 [0.74, 2.21]
Erythromycin (N=32)	Pre DDI	4.69	1.65 [0.69, 3.97]
	DDI exposure	16.04	5.65 [1.41, 22.60]*
	Post DDI	3.48	1.23 [0.46, 3.27]

Note: IR = incidence rate

IRR = incidence rate ratio

* Results are significantly different.

** IRRs were calculated to compare with the non-DDI group IR which is 2.84 per 10 person-year

4.4 Discussion

The drug interactions between statins and CYP3A4 and/or OATP1B1 inhibitors are well recognized. However, studies that focus on the interactions

between simvastatin and specific interacting drugs are very limited. We pre-selected several high-risk drugs that inhibited either CYP3A4 and/or OATP1B1 and conducted a population-based study using administrative claims data to explore the actual exposures and risks of adverse events in CVD patients who were prescribed combination therapy of simvastatin and these pre-specified high risk interacting drugs. Given the well-described risk associated with these DDI exposures, it was expected that there would be very few individuals using these medication combinations. It was also expected that if these combinations were prescribed that there would be a need for greater clinical follow-up to monitor for potential adverse drug events related to DDI exposure. A subgroup analysis was also performed to investigate the interactions between simvastatin and three specific interacting drugs that few previous studies had evaluated and had substantial numbers of DDI-exposures in the study cohort. The study used real world data to provide empirical data on the actual risk of DDI exposure, the length of exposure, and the risk of clinically significant AEs.

The baseline Charlson index score and simvastatin exposure time in the DDI group were significantly higher than the non-DDI group indicating the DDI group was a sicker population with longer simvastatin exposure time. DDI group had more medical follow-up than non-DDI group with a statistically significant result. The AE incidence rate in the DDI group is higher than the non-DDI group but was not statistical significant. Although results are not adjusted for Charlson comorbidity score, we compared the individual comorbidity between DDI and non-DDI group. We found baseline liver and renal disease, which could

potentially affect the AEs during medication exposure, are not statistically different. This may indicate the differences between the two groups are more likely to be explained by DDI. However, the increased comorbidities among those with DDI exposures may also contribute to a greater risk of illness including adverse drug events.

Within the DDI group, the number of physician claims during the DDI exposed time period was more than that during the DDI unexposed time period. This indicates patients were receiving more medical care during the concomitant administration of simvastatin and interacting medications, potentially providing important clinical monitoring or physician visits for AEs. When comparing the pre-DDI and post-DDI time periods, the results showed patients generally had more physician claims in pre-DDI than post-DDI time period. This may be due to patients having more intense physician follow-up in the months after the index CVD event to provide needed medication adjustments and control of cardiovascular risk factors. This study also showed that the number of patients with AEs increased with DDI exposure time which indicates longer DDI exposure may induce higher risk of AEs.

When comparing the three inhibitors, we found clarithromycin and erythromycin had higher risk of AEs than gemfibrozil during DDI therapy. This may be because both clarithromycin and erythromycin are potent inhibitors of CYP3A4 and OATP1B1[121–123], leading to a greater total effect on the simvastatin by influencing two clinically important drug pathways. Gemfibrozil, on the other hand, is primarily dependent on OATP inhibition but not CYP3A4 leading to a

weaker overall inhibition. An interesting finding about gemfibrozil was that it had the highest IR of AEs during pre-DDI time period compared with the other two inhibitors. So we conducted a further investigation and found gemfibrozil group had substantially higher percentage of male patients (72.6%) compared with clarithromycin (55.4%) and erythromycin (46.9%) groups, which indicated the differences may be due to gender differences reflecting potentially different hormonal influences and associated pharmacogenomics. Likewise, the statin dropout rate in gemfibrozil patients was the highest (17%) among all three inhibitors. However, we do not have a solid reason to explain why gemfibrozil has a significant higher Pre-DDI IR. Further studies need to be conducted to see if similar result occurs.

Our findings are consistent with a population-based cohort study[50] which was conducted in Canada showing that the co-prescription of a statin metabolized by CYP3A4 with clarithromycin or erythromycin was associated a higher risk of hospitalization with rhabdomyolysis, acute kidney injury, and all-cause mortality compared with azithromycin. Their outcomes were based on a composite of atorvastatin, simvastatin, and lovastatin and a composite of clarithromycin and erythromycin. Kellick et al.[121] pointed out that gemfibrozil had a less profound effect on the statin medications. In contrast, Mesgarpour et al.[109] found the risk for hospitalization or death in persons receiving clarithromycin is not causally associated with the interactions between statins (atorvastatin, simvastatin, or lovastatin) and clarithromycin. Another cohort study[110] focused on the concomitant use of statins and fibrates indicating that the combination

therapy of simvastatin and gemfibrozil increased the incidence of rhabdomyolysis hospitalizations.

This study had several limitations. Patients may have been misclassified if 1) AEs were not simvastatin-related AEs or induced by other drug interacting factors such as food (grapefruit juice[124–126]) or underlying clinical diseases (liver disease and renal dysfunction). However, the baseline comparisons showed no significant difference between the DDI and non-DDI group; 2) the initial AEs happened before starting interacting drugs and the follow-up visits that occurred after initiating interacting drugs were identified as adverse events due to timing accuracy problems; 3) events were not captured by the medical codes. There may be selection bias since the study population only included subjects with a stable statin dosage, however, it is not clear if this is likely to affect the occurrence of AEs in either the DDI or non-DDI exposed groups, but may affect generalizability of the findings. In addition, IR and physician claim comparison between DDI and non-DDI group were not adjusted for comorbidities. However, we compared individual comorbidities between two groups. Other limitations include the results that were based on the small size group, e.g. erythromycin, were difficult to draw meaningful conclusions, and the associations based on observational study may not be causal.

4.5 Conclusions

The potential statin-drug interactions in CVD patients are common and should be monitored to limit patient adverse events. The risk of adverse events is amplified when concomitant administration of simvastatin with clarithromycin or

erythromycin, which is likely due to the double paths to create drug interactions. The combination of simvastatin and gemfibrozil yield fewer adverse events than the combination with clarithromycin or erythromycin which may be due to single pathway inhibition. If possible, these inhibitors, especially clarithromycin and erythromycin, should be avoided in clinical settings when patients take simvastatin due to the risk of adverse events.

Chapter 5 Prediction for statin-associated adverse events using machine learning technologies

This chapter shows how the component 4 in Figure 1-1 was implemented. Five classic machine learning models were developed to predict one-year risk of AEs among CVD patients after statin initiation.

5.1 Introduction

Machine learning (ML) methods, which are primarily used for prediction and exploratory studies, were used in this project for individualized prediction and data-driven decision making. ML approaches have been increasingly used in healthcare research for prediction because they are preferable to handle the size and complexity of health data and can provide powerful approaches to discover hidden patterns from large datasets. In recent years, the power of ML in diagnosing disease and in predicting treatment outcome empowers physicians and reduces time for decision making in the clinic. Instead of relying on reactive strategies, ML predictive modeling and Big Data approaches were used to develop proactive strategies. ML is intended to support provider decision making to prescribe a personalized statin treatment plan (both statin agent and dosage) based on multiple patient characteristics and concomitant drug therapy to minimize the individual's risk of AEs. In developing a proactive strategy, ML algorithms are used to identify similar patients and their treatment plans in the data, generalize the treatment plan, and predict the personalized treatment plan that minimizes AEs risk to support prescription decisions. In this chapter, I will introduce how classic ML models were developed and how the optimal model

was selected to predict up to one-year risk of AEs among CVD patients after statin initiation. The performance comparison was also investigated among different feature sets including single feature, features selected by forward selection, clinical expert selected features, and all available features. In the future, this optimal predictive model could be integrated into a clinical decision support system to provide personalized statin treatment plans and minimize the risk of AEs.

5.2 Methods

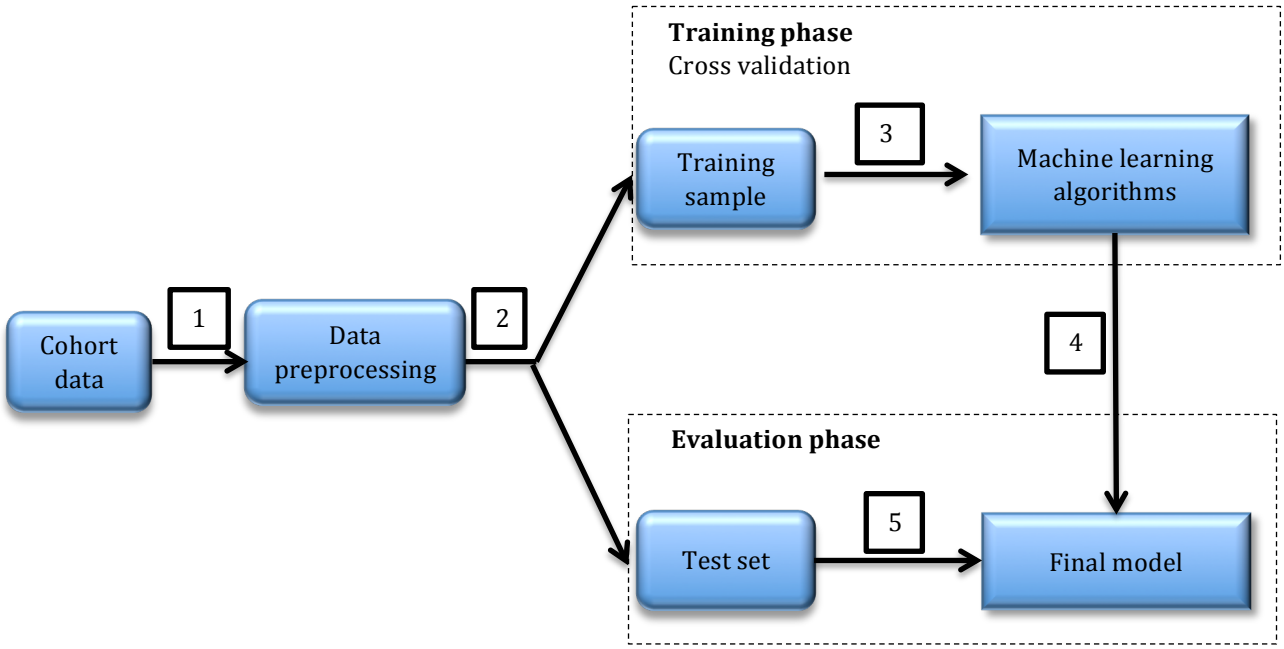


Figure 5- 1 Steps for model development

Figure 5-1 shows all the steps that were implemented to develop machine learning algorithms and select the final optimal model. These include 1) data preprocessing; 2) random separation: divide the cohort patients into training sample and test set; 3) cross validation: validation technique used to train algorithms; 4) final optimal model selection; 5) evaluation: apply final model on

the test set. Matlab version R2016b was used as the software to perform the ML tasks.

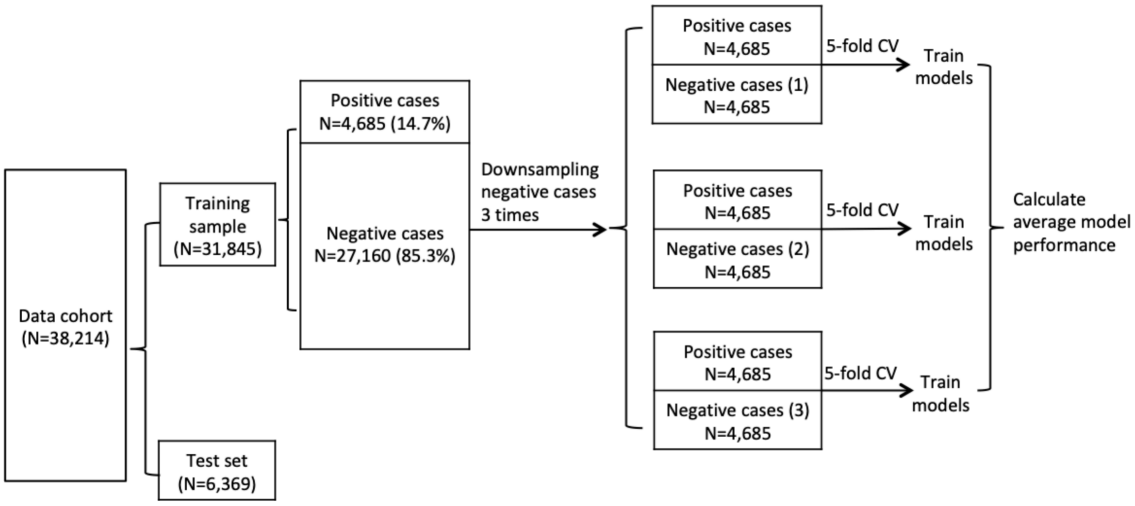
5.2.1 Data preprocessing

Variables that were skewed and had a wide distributional range were log transformed to obtain a normal distribution[127]. In addition, some continuous variables that were measured in different scales were not directly comparable and thus did not contribute equally to prediction. For instance, medical costs can range across multiple orders of magnitude causing cost data to have more weight than smaller value variables such as statin dosage. To reduce the chance that these variables dominate the model, feature scaling was used to rescale all continuous values into the range [0,1]. The formula is shown below, where X is the original value; $\mathrm{X_{min}}$ and $\mathrm{X_{max}}$ are the minimum and the maximum value of X. The new value X' was obtained after the normalization.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The cohort was divided into two datasets: the training sample and test set. The original dataset was randomly divided into six equal size subsets. One of the subsets was randomly selected as the test set, which was used to evaluate whether the final optimal model can make a reliable prediction on unseen data. Among the remainder of five subsets (training sample), only 14.7% patients were positive cases (patients who had any type of AEs, minority class) causing an imbalanced dataset. Models built on class-imbalanced data are often biased towards the majority class and cannot accurately predict the minority class. To balance the data, I randomly downsampled the majority class (patients who did

not have any AEs) to obtain an equal size of the majority class and the minority class (see Figure 5-2). In addition, the downsampling process was repeated three times with replacement to reduce the selection bias.



Note: Positive cases were patients who had adverse events.

Negative cases were patients who did not have any adverse events.

CV = cross validation.

Figure 5- 2 Process of downsampling

5.2.2 Predictive models development

Five classic ML methods were initially investigated, including generalized linear model (GLM), support vector machine (SVM), decision tree, random forest, and artificial neural network (ANN). Final model selection was based on model performance and model complexity.

GLM models including linear, interactions, pure quadratic, and quadratic functions were investigated. For SVM, kernel function is one of the hyperparameters that need to be tuned. Three kernel functions, including linear kernel, RBF kernel, and polynomial kernel, were tested. Three polynomial kernel

orders (2, 3, and 4) were also tested. Another hyperparameter is called box constraint (BC), which is a cost/penalty to the misclassification to adjust the boundary and the number of support vectors. When the data is not perfectly separable, the training algorithm must allow some misclassification in the training set. In this case, it is applying a cost to the misclassification. The higher the box constraint, the higher the cost of the misclassified points leading to a more strict separation of the data and more conservative classification. Seven BC numbers were tested: 0.001, 0.01, 0.1, 1, 10, 100, and 1000, which covered a wide range. Gini and entropy split metrics were investigated when developing the decision tree model[128]. Ten numbers (2^1, 2^2, 2^3, continuing all the way to 2^{10}) were chosen for both the maximal number of splits and the minimum number of leaf node observations. For random forest, two hyperparameters, including tree number (the maximum number of trees to build) and feature number (the number of features a random forest selects when splitting a node), were tuned. Tree number was tested from 1 to 500. The default value of feature number is $log_2(M+1)$, where M is the total number of features [129]. However, when many of the variables are categorical, the number of features must be increased to about 2 to 3 times of the default value in order to get enough strength to provide good accuracy on test set[129]. Since a total of 35 features were included, the default NumFeature is 5.2. In addition, many of the predictors are categorical, thus the number of features need to be increase to 10-16. Therefore, seven feature numbers were tested including 5, 7, 9, 11, 13, 15 and 17. For the neural network classification, three common training functions including Levenberg-Marquardt

optimization (LM), scaled conjugate gradient (SCG), and Bayesian regularization (BR) were tested. Previous studies[130,131] showed that 1 or 2 hidden layers should solve most of the problems. Therefore, we tested hidden layer size from 1 to 5.

A 5-fold cross validation was performed during model development. AUC ROC was used as the primary evaluation metric. In addition, sensitivity was also an important factor to measure model performance, because we did not want to misclassify the true positive cases. Computational complexity was also considered when models had similar AUC ROC and sensitivity. A simple model with low computational complexity was better than a complex model, because simple model is quicker to build, easier to implement, and easier to interpret than complex model.

5.2.3 Outcome

A combination of all types of AEs (myopathy, rhabdomyolysis, renal damage, liver damage, and statin poisoning events) was used as the predicted outcome. Positive cases who had any kinds of AEs were coded as 'AE=1'.

5.3 Results

38,214 patients were included in our cohort population. 6,369 patients were randomly selected as the test set. Three downsampled training datasets were created from the original training dataset. Each training set contained 9,370 patients which included 4,685 positive cases who had at least one type of AEs and 4,685 negative cases sampled from patients who did not have any type of AEs.

5.3.1 Performance of different feature sets

The original feature set included 74 features including age, gender, comorbidities, medical cost, statin cost, and other variables (See details in

Supplementary Table 3). Two comorbidity scoring methods, Charlson and Elixhauser comorbidities, are included in the original feature set. Only one of the comorbidity scores will be selected as a feature based on their predictive performances. Individual comorbidities were also selected by automated forward feature selection.

GLM was used to evaluate and compare the prediction ability among four feature sets: single feature, 35 features selected by GLM forward selection (Supplementary Table 5), 32 domain expert selected features (Supplementary Table 5), and 74 all features (Supplementary Table 3). These comparisons showed: 1) how much improvement when using more features compared with the single feature; 2) the differences between machine selected features and clinical expert selected features based on domain knowledge. Table 5-1 shows the comparisons among three feature sets. The first column lists the top 10 single features with the highest AUC ROC (see details in Supplementary Table 4 for all the ordered single features). Each feature was used as the only one predictor in GLM and evaluated separately. AUC ROC was used to evaluate model performance and rank features. The second column lists the top 10 predictors selected by GLM forward selection (see details in Supplementary Table 5). The last column shows whether these features were considered by a domain expert. This table shows that the features selected by GLM forward selection were clinically meaningful since they are highly consistent with what the clinical expert selected.

Table 5- 1 Features selected by different methods

Features	Single feature (Top 10)	GLM forward selection features (Top 10)	Clinical expert selected features
Age	X	X	X
Gender	X	X	X
Charlson comorbidity score	X	X	X
Statin strength	X	X	X
Statin initiation gap[1]	X	X	X
Statin exposure time	X	X	X
Health plan type[2]	X	X	X
Statin cost (out-of-pocket)[3]	X		X
Initial 30-day medical cost	X	X	X
Prescriber specialty		X	X
Elixhauser comorbidity score	X		
Fluid and electrolyte disorders		X	X

Note: 1. Statin initiation gap is defined as the gap between CVD index date and statin initiation date.

2. Health plan type includes commercial and Medicare Advantage.

3. Cost was calculated as cost per 30 days.

Figure 5-3 shows the performance comparison among different feature sets. The top three single predictors with best performance were included for illustration purposes. When compared with other feature sets, the results indicated that a single predictor did not have enough prediction power. Adding

more features improved the performance. Features selected by GLM forward selection had the highest AUC ROC and sensitivity. Clinical expert selected features had similar performance with GLM selected features. All features had relatively low performance compared with the other feature sets which indicates it may included some irrelevant features. The 35 features selected by GLM forward selection were used in the subsequent models (Supplementary Table 5).

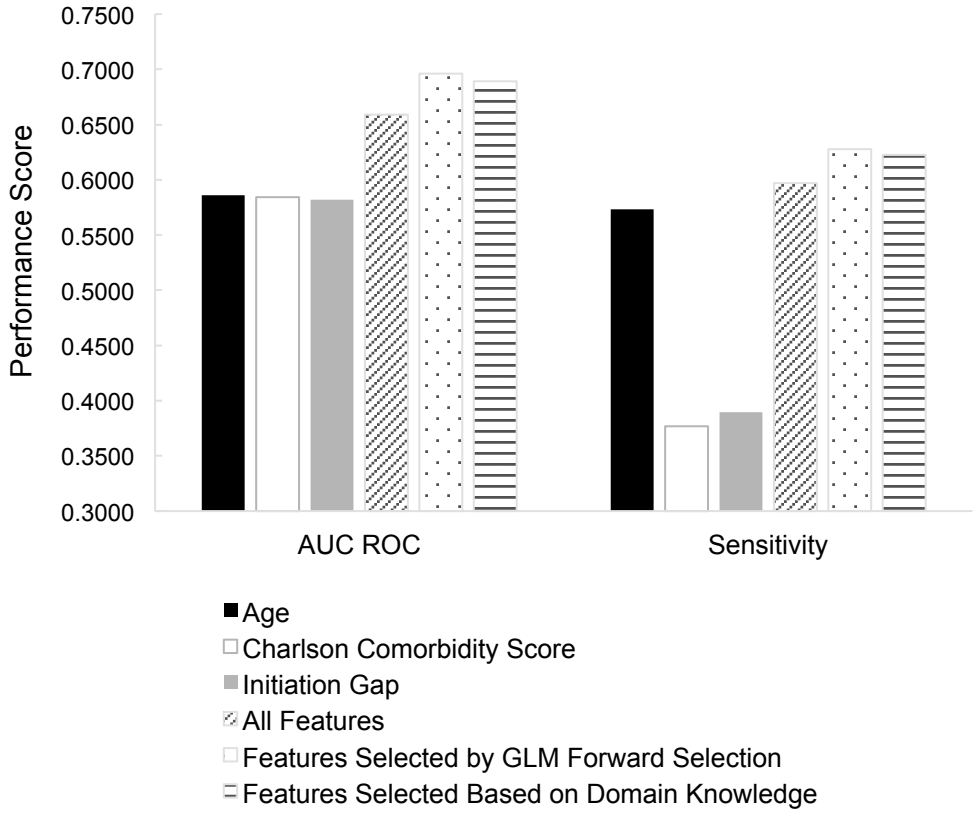


Figure 5- 3 Performances comparison among different feature sets

5.3.2 Baseline characteristics comparison between randomized downsampling datasets

The three downsampling datasets were randomly selected. Baseline characteristics comparison among the three groups – ANOVA analysis, were

performed to verify if the random process was successfully conducted. Table 5-2 shows that the patient baseline characteristics among the three groups are very similar. The randomized process was completed successfully.

Table 5- 2 Baseline characteristics comparison among the three downsampling groups.

Patient characteristics	Sample 1, (N=9,370)	Sample 2, (N=9,370)	Sample 3, (N=9,370)	p-value
Statin exposure time (days)	221.3	220.5	222.8	0.528
Age (years)	65.5	65.5	65.4	0.965
Male (%)	57.0	56.3	56.9	0.577
Charlson index score	1.27	1.28	1.28	0.933

5.3.3 Models development and comparison

Model performance will be considered to be good if AUC ROC is greater than 0.80, moderate if values are between 0.60 and 0.80, and poor if values are less than 0.60.

For GLM model, four models were included (see Table 5-3). The model performances are shown in Figure 5-4. Model 1 (linear model) had the highest AUC ROC. Its sensitivity is a little bit lower than the pure quadratic model, however, it is the simplest model among all models. Thus, the GLM linear model was identified as the best GLM model.

Table 5- 3 GLM Models

Model	Model specific	Distribution

Model 1	Linear	Binomial
Model 2	Interaction	Binomial
Model 3	Pure quadratic	Binomial
Model 4	Quadratic	Binomial

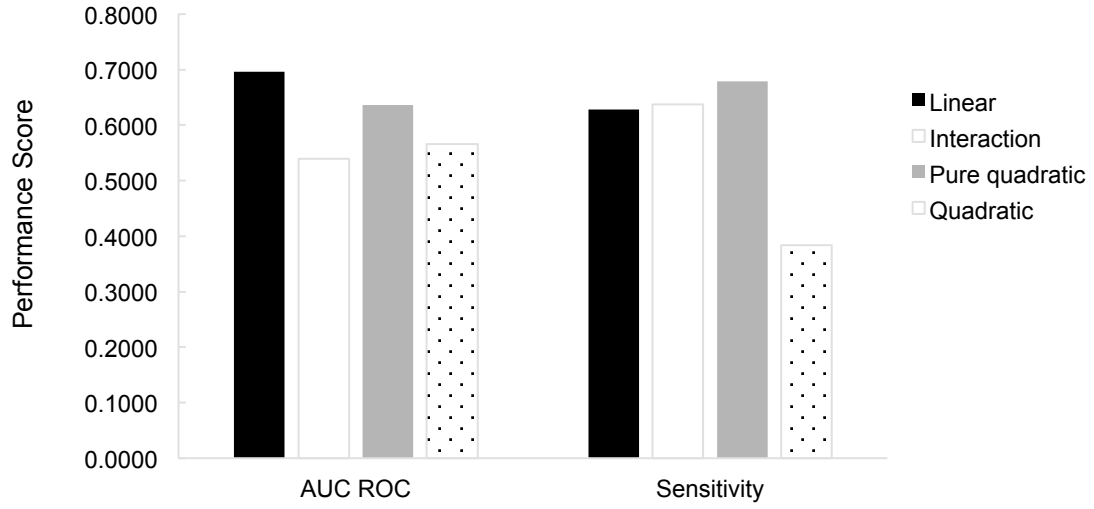
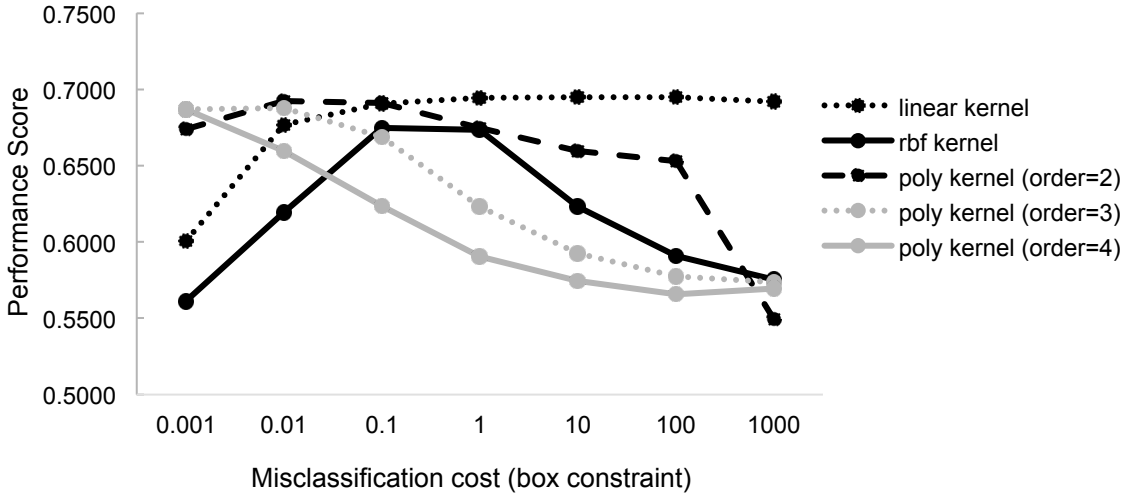


Figure 5- 4 Performances comparison of 4 different GLM models

Thirty-five SVM models were built and compared. AUC ROC and sensitivity comparisons are shown in Figure 5-5 and Figure 5-6. AUC ROC and sensitivity generally increased as the box constraint increased, and then stopped increasing or decreased after a specific box-constraint value. Although the model with the radial basis function kernel and box constraint = 0.01 had the highest sensitivity, it had relatively low AUC ROC. Thus, we selected the linear kernel with box constraint=0.1 as the best SVM model since both AUC ROC and sensitivity were relatively high.

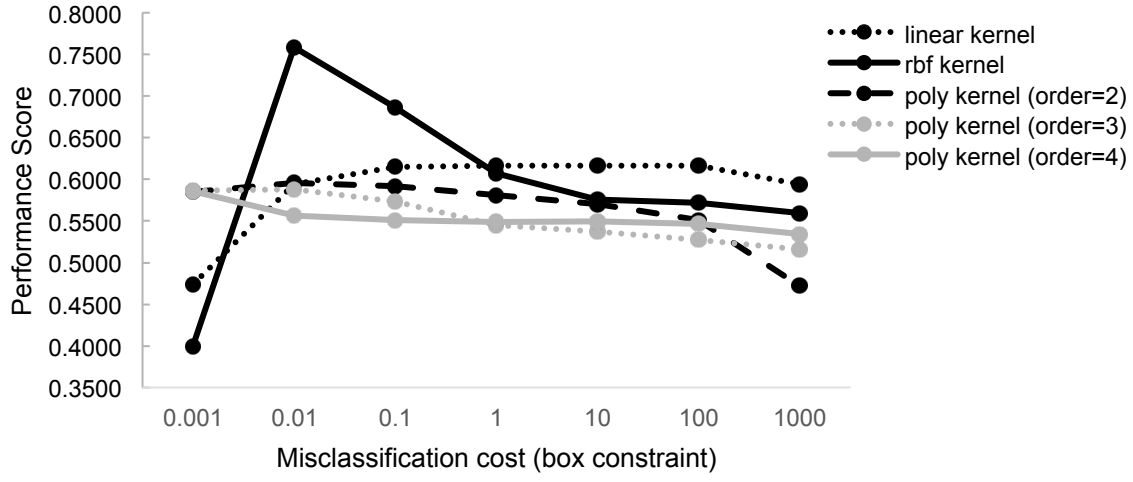


Note: bc = box constraint

rbf = radial basis function

poly kernel = polynomial kernel

Figure 5- 5 AUC ROC of different SVM models



Note: bc = box constraint

rbf = radial basis function

poly kernel = polynomial kernel

Figure 5- 6 Sensitivity of different SVM models

The random forest classifier was applied using classification method. The tree number was determined using out-of-bag (OOB) classification error[132] which is an error estimation technique often used to evaluate the accuracy of a random forest and determine the optimal number of trees. Lower number of trees results less complex model. Figure 5-7 shows the OOB error decreased as the tree number increased. The OOB error tended to be stable after it reached a certain number. As a result, 250 trees were selected as the maximum tree number because the out-of-bag classification error was relatively stable after reaching 250. To select the optimal features to split a node, 7 numbers were tested which are shown in Figure 5-8. We selected 13 as the optimal feature number because it had the highest AUC ROC and sensitivity.

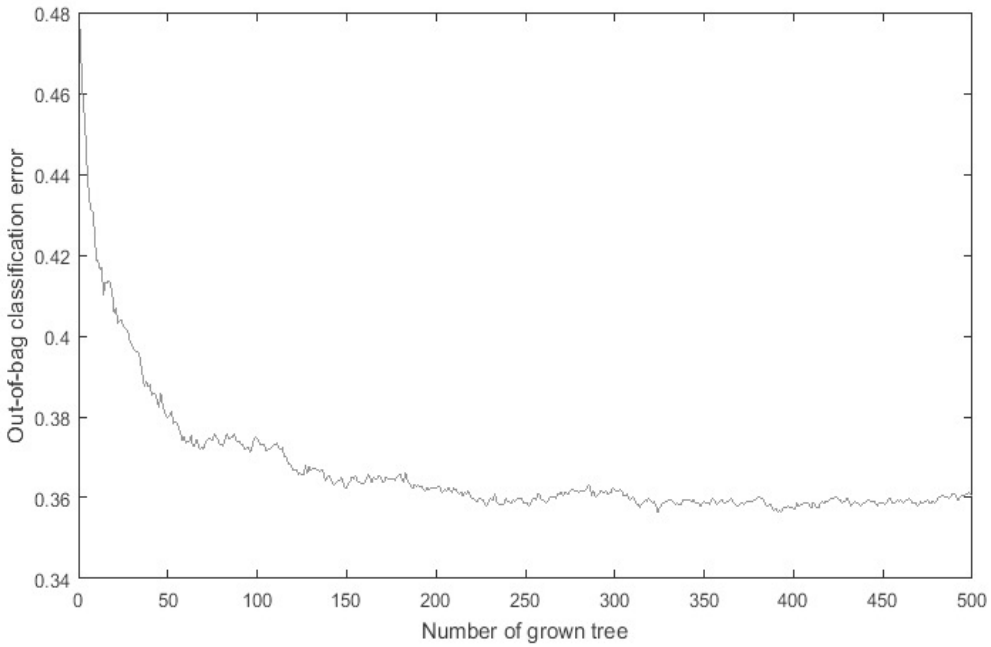


Figure 5- 7 Out-of-bag error relationship to tree number

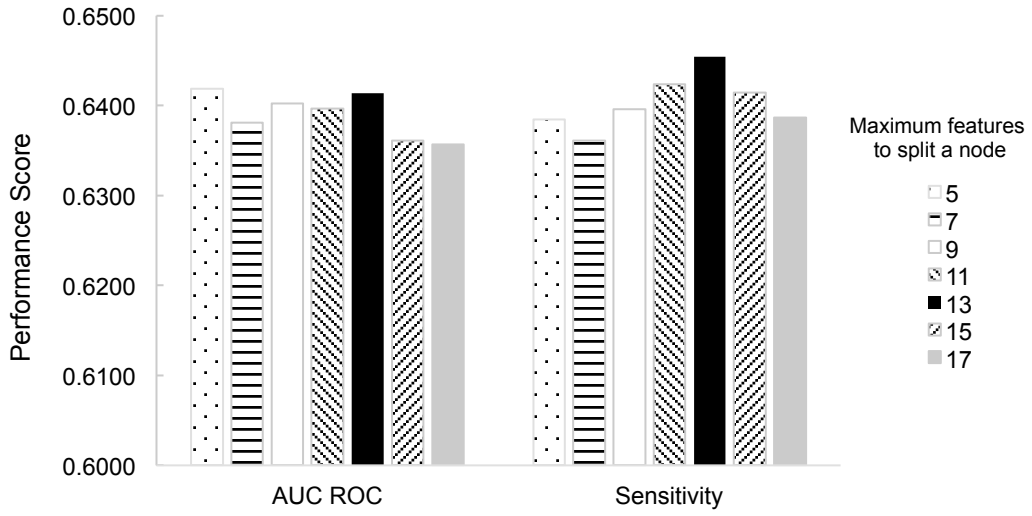


Figure 5- 8 Performances of random forest

For decision tree analysis, two split metrics were tested: Gini impurity and entropy. As discussed in 5.2.2, ten numbers (2^1, 2^2, 2^3, continuing all the way to 2^{10}) were chosen for both the maximal number of splits and the minimum number of leaf node observations. Thus, for each split metric, 100 models were tested. The AUC ROC and sensitivity comparisons for the two split metrics are shown in Figure 5-9, Figure 5-10, Figure 5-11, and Figure 5-12. The x-axis represents the minimum number of leaf node observations (2^{bb}) and the y-axis represents performance score (AUC ROC or sensitivity). The legend represents the maximal number of splits (2^{aa}). The performances of Gini impurity and entropy were very similar. Gini split metric was selected instead of entropy because entropy requires computation of logarithmic functions, which is more computationally intensive than Gini impurity. Gini model with maximum number of split = 512 and minimum number of leaf node observation =128 was chosen after considering AUC ROC, sensitivity, and model complexity.

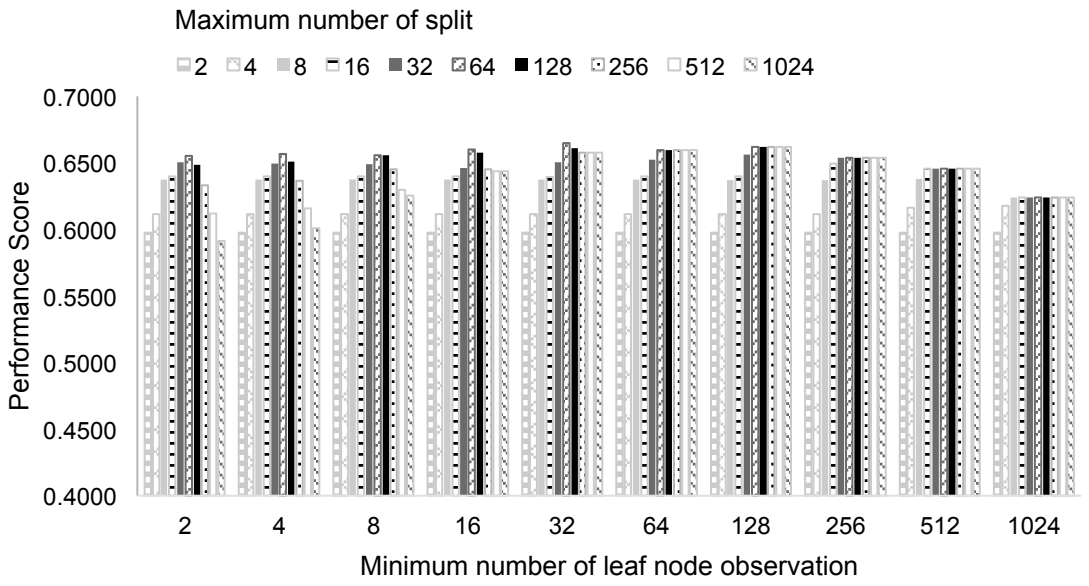


Figure 5- 9 AUC ROC of different decision tree models (Gini)

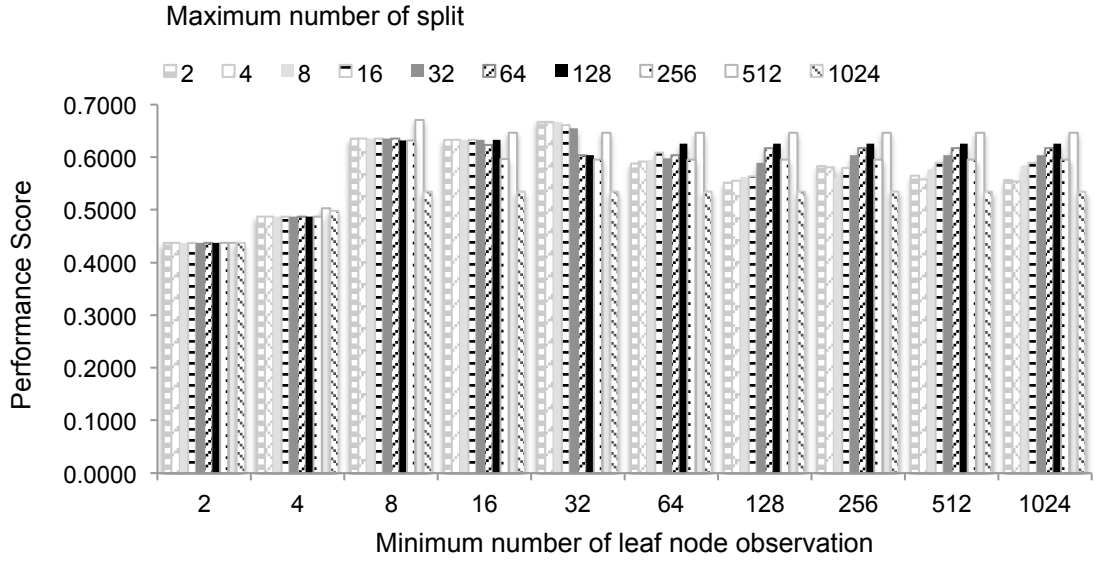


Figure 5- 10 Sensitivity of different decision tree models (Gini)

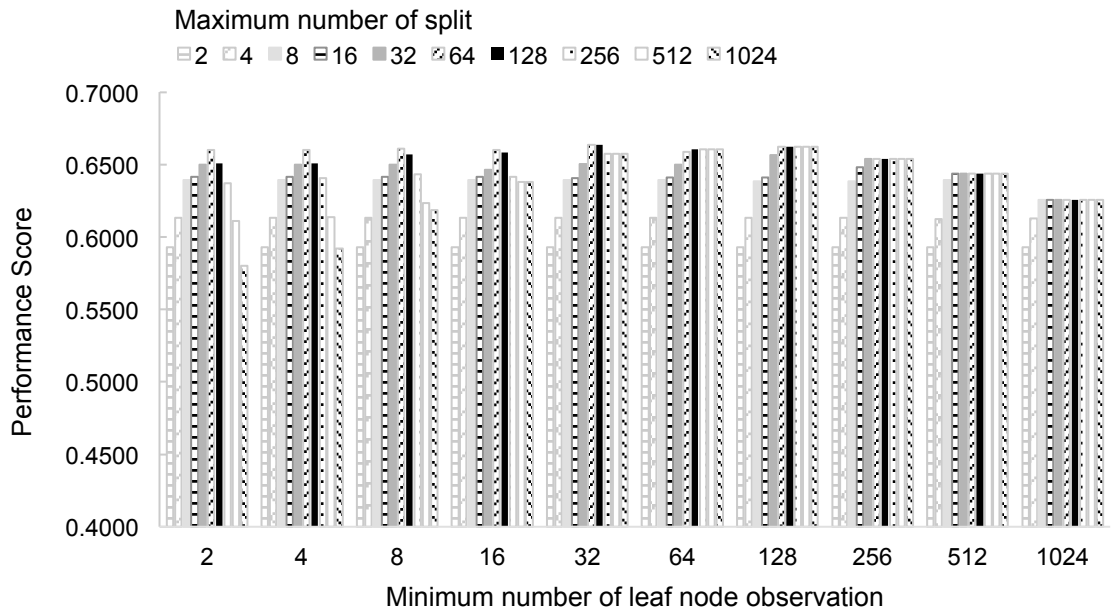


Figure 5- 11 AUC ROC of different decision tree models (Entropy)

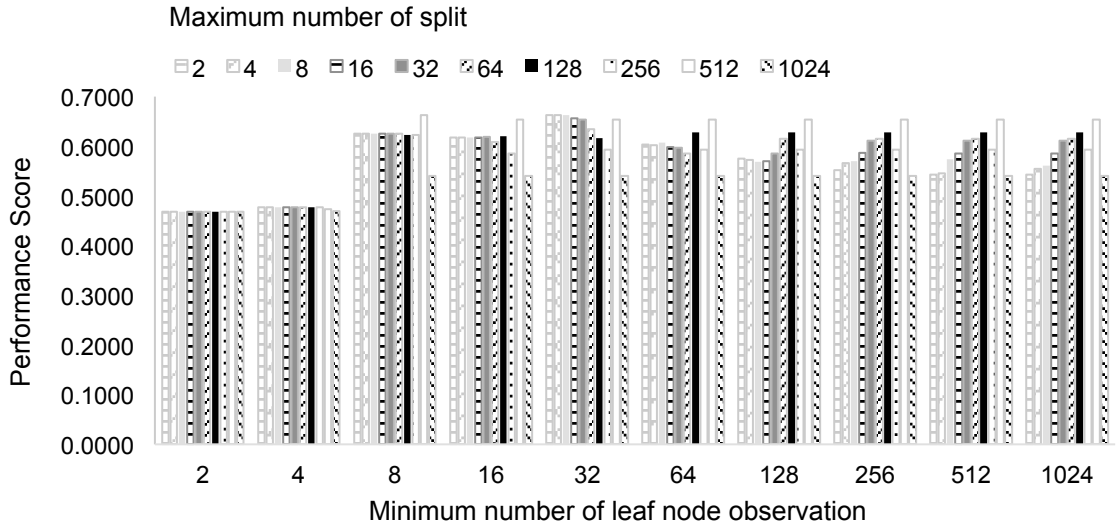


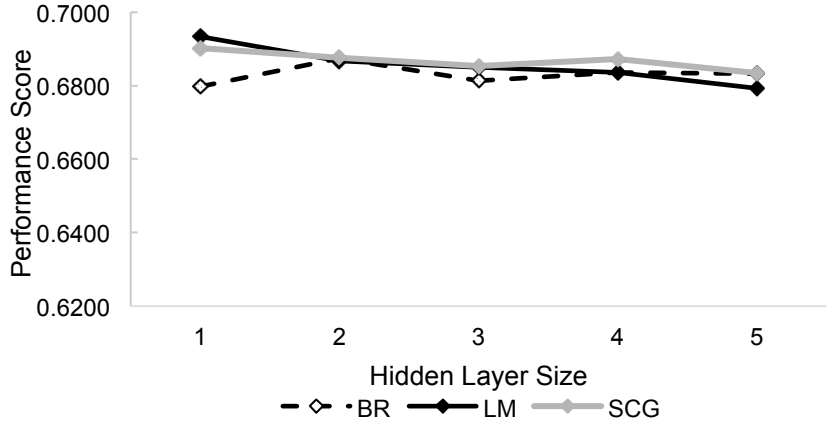
Figure 5- 12 Sensitivity of different decision tree models (Entropy)

For the ANN classifier, five hidden layer sizes (from 1 to 5) and three network training functions were tested, resulting in a total of 15 models. Performance comparisons are shown in Figure 5-13. Levenberg-Marquard training function with a hidden layer size=1 was selected as the best ANN model

because it had the highest AUC ROC. Although its sensitivity was a little bit lower than the Bayesian regularization training function with hidden layer size=5, the latter model is more complicated.

To select the final optimal model, I compared the performance of the following classifiers: 1) GLM: linear model with binomial distribution; 2) SVM: linear kernel with box constraint=0.1; 3) random forest: maximum 250 trees and 13 candidate predictors randomly drawn for a split; 4) decision tree: Gini model with maximum number of split = 512 and minimum number of leaf node observation =128; and 5) ANN: Levenberg-Marquard training function with hidden layer size=1. Figure 5-14 shows that GLM, SVM and ANN had relatively higher AUC ROC than the other two classifiers. However, GLM model was selected as the final optimal model as it has the highest sensitivity and is the simplest model among all models.

(a)



(b)

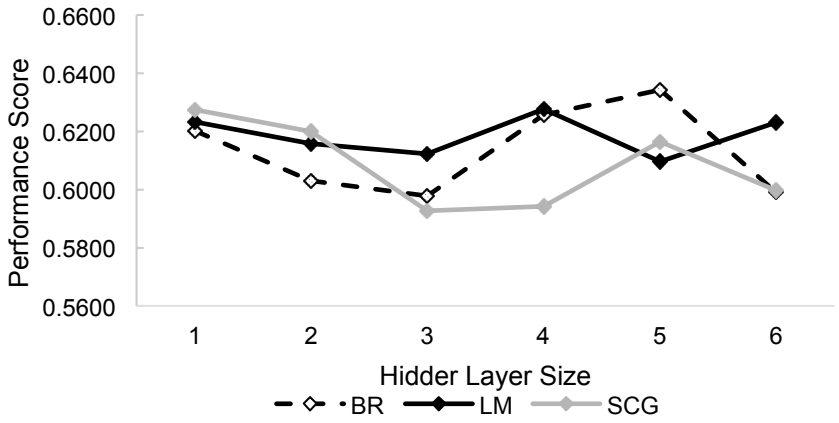


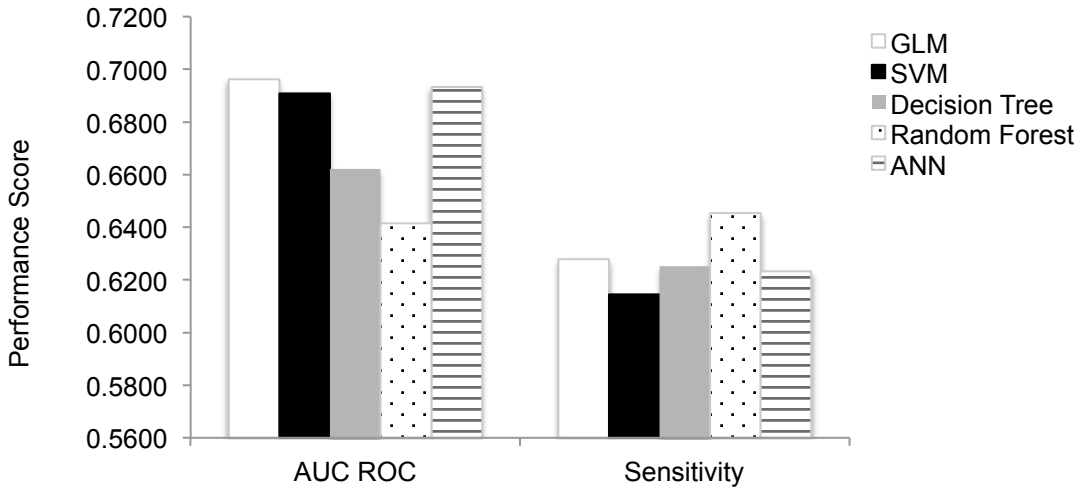
Figure 5- 13 (a) AUC ROC and (b) sensitivity of different artificial neural network

Note: BR = Bayesian regularization

LM = Levenberg-Marquard

SCG = scaled conjugate gradient

Finally, the GLM model was validated on the test data, which was never used when built models, to assess model generalizability. The AUC ROC of the test dataset was 0.70.



Note: GLM = generalized linear model

SVM = support vector machine

ANN = artificial neural network

Figure 5- 14 Performances of different classification methods

5.4 Discussion

In this chapter, a machine learning model was developed using OLDW claims data to predict the one-year risk of adverse events after prescribing a statin treatment plan. This predictive model could potentially improve the adherence of statins and reduce the risk of statin-associated AEs. It can also be used in a future clinical decision support system by automatically comparing AE risks among different statin treatment plans and selecting the best statin treatment plan for individual patient. The results demonstrated a moderate capacity to predict the one-year statin-associated adverse events.

There are several study limitations. First, some of the clinically important predictors were not included due to representation gaps in the claims data, such as lab data, BMI, and smoking status. BMI [133] and smoking status [134,135]are clinically important when predicting statin AEs. However, only 10% of the cohort population had this information in claims data. Lab data was excluded from the analysis for several reasons: 1) only 30% to 40% of people have any lab results, however, the presence of a lab result doesn't ensure the patients' lab records are complete. In this cohort population, only about 20% patients had lab results. For some lab values, such as the creatinine kinase level, less than 1% patients had test results; 2) only contains outpatient lab results and most common sample types are serum, urine, and blood based tests ; 3) not all outpatient lab test results are available; only results that are provided by specific clinical laboratories. Therefore, lab results processed outside of these specific clinical

laboratories are not available. In addition, this model only focused on older population (>40 years of age). Furthermore, our models were built based on combination of all types of AEs. Models for specific AEs have not yet been investigated since some of the AEs were very rare. In our cohort population, only 43 patients had rhabdomyolysis and 24 patients had poisoning events. It's not possible to build adequate models based on such small sample size.

5.5 Conclusion

Machine learning technologies were applied to develop predictive models to predict statin adverse events in CVD patients. In this feasibility study, the predictive model had a moderate performance with an AUC ROC of 0.70. This model has the potential to be applied to a clinical decision support system to help improve statin adherence and minimize the risk of severe statin-associated symptoms.

Chapter 6 Summary and future directions

Statin associated adverse events is the primary reason for statin discontinuation. The consequence is that important clinical and treatment benefits of statin use for lowering plasma cholesterol levels (and therefore risk of primary and secondary cardiovascular events) are frequently lost due to discontinuance of statin treatment following occurrence of AEs, leading to increased cholesterol levels and increased cardiovascular events. This dissertation investigated statin associated adverse events in three aspects: AEs risk evaluation in CVD patients who initiated statin agent for secondary prevention, simvastatin drug-drug interactions, and adverse events prediction. OLDW claims data was used as the data source to do the secondary data analysis which is a growing trend and is an important resource for population-based healthcare research.

The first part demonstrated how the cohort population was extracted and how this cohort was used to investigate specific adverse events in all CVD statin users. The descriptive analysis of demographic and other characteristics were done for each statin agent group and each statin tolerance group (continuous, discontinued, and dropout). Incidence rates were calculated for each adverse event stratified by statin agents and statin tolerance groups.

The second part investigated drug-drug interactions focused on simvastatin. Comparisons between DDI group and Non-DDI group and among the three time periods within DDI group were conducted. This study analyzed the effect of CYP3A4 and/or OATP1B1 inhibitors on simvastatin.

The last part discussed how the predictive models were developed to predict the up to one-year risk of statin-associated adverse events for CVD patients. The performances were compared among five machine learning algorithms. SVM was selected as the final optimal model which had a moderate predictive ability with an AUC ROC of 0.69 on a set of unseen data.

In the future, adult patients from 18 to 40 years of age should be included to increase the generalizability of the results of the predictive model,. Other large-scale databases may be considered as well. If possible, the combination of multiple healthcare databases is also a good way to increase the sample size, obtain a wider range of population, and make the results more generalizable.

Model accuracy is expected to be further improved. Several ways can be considered including adding more years of data and more potentially important predictors (e.g., lab data and BMI) by linking with the EHR data. In addition, deep learning, as a subset of machine learning, becomes more and more popular because it outperforms other methods in many domains. Most of the deep learning models are based on the artificial neural networks. Feature selection is embedded in the learning process, so domain expertise is less needed and irrelevant variables will have very small impacts on the prediction. Another important reason to use deep learning is that complex interactions among multiple variables are naturally included in the prediction model (consequently, the model predicts using many interactive layers of nodes and functions). Using deep learning techniques on our dataset may achieve higher accuracy since they

have improved classification ability and model accuracy. More outcomes will be considered in the future study. A deep-learning model will predict both statin associated adverse events and statin discontinuation. A more functional platform will be developed to identify the optimal statin treatment plans that not only reduce risk of adverse events, but also optimize LDL reduction and reduce statin discontinuation for a given patient profile. This platform could be potentially applied to a clinical decision support system to provide personalized statin treatment plans, minimize the risk of adverse events, and potentially improve statin adherence. Knowing which patients are at higher risk of adverse effects of statin therapy is valuable. The patient-specific estimated risk could be displayed in the EHR within the normal workflow of a clinic or hospital. An alert generated by a clinical decision support application in an EHR could assist healthcare providers to identify high-risk patients at the point of care, efficiently make a decision on appropriate therapy, and prompt a nursing protocol that included patient education, follow-up phone calls, and more frequent clinic visits during the statin therapy period. Any patient at a certain threshold of risk could be automatically placed on the protocol to improve early identification of problems for potential interventions.

BIBLIOGRAPHY

1. Jankel CA, Fitterman LK. Epidemiology of Drug-Drug Interactions as a Cause of Hospital Admissions. Drug Saf. 1993;9:51–9.
2. Baigent C, Blackwell L, Emberson J, et al. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. Lancet. 2010;376:1670–81.
3. Delahoy PJ, Magliano DJ, Webb K, Grobler M, Liew D. The relationship between reduction in low-density lipoprotein cholesterol by statins and reduction in risk of cardiovascular outcomes: An updated meta-analysis. Clin Ther. 2009;31:236–44.
4. Cohen JD, Brinton EA, Ito MK, Jacobson TA. Understanding Statin Use in America and Gaps in Patient Education (USAGE): An internet-based survey of 10,138 current and former statin users. J Clin Lipidol. 2012;6:208–15.
5. Casas JP, Taylor FC, Ward KJ, et al. Statins for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2013;
6. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129:S1–45.
7. Brugts JJ, Yetgin T, Hoeks SE, et al. The benefits of statins in people without established cardiovascular disease but with cardiovascular risk factors: meta-analysis of randomised controlled trials. BMJ. 2009;338:b2376.
8. Schwartz GG, Steg PG, Szarek M, et al. Alirocumab and cardiovascular outcomes after acute coronary syndrome. N Engl J Med. 2018;379:2097–107.
9. National Center for Health Statistics. Health, United States, 2013: With Special Feature on Prescription Drugs. Hyattsville, MD. 2014.
10. Rublee DA, Chen S-Y, Mardekian J, Wu N, Rao P, Boulanger L. Evaluation of Cardiovascular Morbidity Associated with Adherence to Atorvastatin Therapy. Am J Ther. 2012;19:24–32.
11. Mancini GBJ, Baker S, Bergeron J, et al. Diagnosis, Prevention, and Management of Statin Adverse Effects and Intolerance: Proceedings of a Canadian Working Group Consensus Conference. Can J Cardiol. 2011;27:635–62.
12. Maningat P, Gordon BR, Breslow JL. How do we improve patient compliance and adherence to long-term statin therapy? Curr Atheroscler

Rep. 2013;15:291.

13. Silva MA, Swanson AC, Gandhi PJ, Tataronis GR. Statin-related adverse events: A meta-analysis. Clin Ther. 2006;28:26–35.

14. Pohjola-Sintonen S, Julkunen H. Muscle-related adverse effects of statins. Duodecim. 2014;130:1622–7.

15. Golomb BA, Evans MA. Statin adverse effects : a review of the literature and evidence for a mitochondrial mechanism. Am J Cardiovasc Drugs. 2008;8:373–418.

16. Tomaszewski M, Stępień KM, Tomaszewska J, Czuczwar SJ. Statin-induced myopathies. Pharmacol Reports. 2011;63:859–66.

17. Hippisley-Cox J, Coupland C. Unintended effects of statins in men and women in England and Wales: Population based cohort study using the QResearch database. BMJ. 2010;340:c2197.

18. Sakamoto K, Kimura J. Mechanism of statin-induced rhabdomyolysis. J Pharmacol Sci. 2013;123:289–94.

19. Björnsson ES. Drug-induced liver injury: an overview over the most critical compounds. Arch Toxicol. 2015;89:327–34.

20. Kon RH, Russo MW, Ory B, Mendys P, Simpson RJ. Misperception among physicians and patients regarding the risks and benefits of statin treatment: the potential role of direct-to-consumer advertising. J Clin Lipidol. 2008;2:51–7.

21. Silva M, Matthews ML, Jarvis C, et al. Meta-analysis of drug-induced adverse events associated with intensive-dose statin therapy. Clin Ther. 2007;29:253–60.

22. Chu PH, Chen WJ, Chiang CW, Lee YS. Rhabdomyolysis, acute renal failure and hepatopathy induced by lovastatin monotherapy. Jpn Heart J. 1997;38:541–5.

23. Corpier CL, Jones PH, Suki WN, et al. Rhabdomyolysis and Renal Injury With Lovastatin Use: Report of Two Cases in Cardiac Transplant Recipients. JAMA J Am Med Assoc. 1988;260:239–41.

24. Hansen KE, Hildebrand JP, Ferguson EE, Stein JH. Outcomes in 45 patients with statin-associated myopathy. Arch Intern Med. 2005;165:2671–6.

25. U.S. Food and Drug Administration. Fda Expands Advice On Statin Risks [Internet]. 2011.

26. Golomb BA, McGraw JJ, Evans MA, Dimsdale JE. Physician response to patient reports of adverse drug effects: Implications for patient-targeted adverse effect surveillance. Drug Saf. 2007;30:669–75.

27. Zhang H, Plutzky J, Skentzos S, et al. Discontinuation of statins in routine care settings: a cohort study. Ann Intern Med. 2013;158:526–34.

28. Heeschen C, Hamm CW, Laufs U, Snapinn S, Böhm M, White HD. Withdrawal of statins increases event rates in patients with acute coronary syndromes. Circulation. 2002;105:1446–52.

29. Bitzur R, Cohen H, Kamari Y, Harats D. Intolerance to statins: Mechanisms and management. Diabetes Care. 2013;36:S325-30.

30. Bruckert E, Hayem G, Dejager S, Yau C, Bégaud B. Mild to moderate muscular symptoms with high-dosage statin therapy in hyperlipidemic patients - The PRIMO study. Cardiovasc Drugs Ther. 2005;19:403–14.

31. Golomb BA, Evans MA. Statin adverse effects: A review of the literature and evidence for a mitochondrial mechanism. Am J Cardiovasc Drugs. 2008;8:373–418.

32. Sikka P, Kapoor S, Bindra VK, Sharma M, Vishwakarma P, Saxena KK. Statin intolerance: now a solved problem. J Postgrad Med. 2011;57:321–8.

33. Brewer HB. Benefit-risk assessment of Rosuvastatin 10 to 40 milligrams. Am J Cardiol. 2003;92:23–9.

34. Fernandez G, Spatz ES, Jablecki C, Phillips PS. Statin myopathy: A common dilemma not reflected in clinical trials. Vol. 78, Cleveland Clinic Journal of Medicine. 2011. page 393–403.

35. RxFiles, 11th Edition.

36. Deng JW, Song IS, Shin HJ, et al. The effect of SLCO1B1*15 on the disposition of pravastatin and pitavastatin is substrate dependent: The contribution of transporting activity changes by SLCO1B1*15. Pharmacogenet Genomics. 2008;18:424–33.

37. Ho RH, Tirona RG, Leake BF, et al. Drug and Bile Acid Transporters in Rosuvastatin Hepatic Uptake: Function, Expression, and Pharmacogenetics. Gastroenterology. 2006;130:1793–806.

38. Choi HY, Bae KS, Cho SH, et al. Impact of CYP2D6, CYP3A5, CYP2C19, CYP2A6, SLCO1B1, ABCB1, and ABCG2 gene polymorphisms on the pharmacokinetics of simvastatin and simvastatin acid. Pharmacogenet Genomics. 2015;25:595–608.

39. Niemi M. Transporter pharmacogenetics and statin toxicity. Clin Pharmacol Ther. 2010;87:130–3.

40. Kalliokoski A, Niemi M. Impact of OATP transporters on pharmacokinetics. Br J Pharmacol. 2009;158:693–705.

41. Kopplow K. Human Hepatobiliary Transport of Organic Anions Analyzed by Quadruple-Transfected Cells. Mol Pharmacol. 2005;68:1031–8.

42. Fujino H, Saito T, Ogawa S, Kojima J. Transporter-mediated influx and efflux mechanisms of pitavastatin, a new inhibitor of HMG-CoA reductase. J Pharm Pharmacol. 2005;57:1305–11.

43. Bailey DG, Dresser GK. Interactions between grapefruit juice and cardiovascular drugs. Am J Cardiovasc Drugs. 2004;4:281–97.

44. Dresser GK, Bailey DG, Leake BF, et al. Fruit juices inhibit organic anion transporting polypeptide-mediated drug uptake to decrease the oral availability of fexofenadine. Clin Pharmacol Ther. 2002;71:11–20.

45. Ming EE, Davidson MH, Gandhi SK, et al. Concomitant use of statins and CYP3A4 inhibitors in administrative claims and electronic medical records databases. J Clin Lipidol. 2008;2:453–63.

46. Toth PP, Farnier M, Tomassini JE, Foody JM, Tershakovec AM. Statin combination therapy and cardiovascular risk reduction. Future Cardiol. 2016;12:289–315.

47. Bottorff MB. Statin safety and drug interactions: clinical implications. Am J Cardiol. 2006;97:S27–31.

48. Morival C, Westerlynck R, Bouzillé G, Cuggia M, Le Corre P. Prevalence and nature of statin drug-drug interactions in a university hospital by electronic health record mining. Eur J Clin Pharmacol. 2018;74:525–34.

49. Kellick KA, Bottorff M, Toth PP. A clinician's guide to statin drug-drug interactions. J Clin Lipidol. 2014;8:S30-46.

50. Patel AM, Shariff S, Bailey DG, et al. Statin toxicity from macrolide antibiotic coprescription. Ann Intern Med. 2013;158:869–76.

51. Molden E, Andersson KS. Simvastatin-associated rhabdomyolysis after coadministration of macrolide antibiotics in two patients. Pharmacotherapy. 2007;27:603–7.

52. Wagner J, Suessmair C, Pfister HW. Rhabdomyolysis caused by co-medication with simvastatin and clarithromycin. J Neurol. 2009;256:1182–3.

53. Zhanel GG, Hisanaga T, Wierzbowski A, Hoban DJ. Telithromycin in the treatment of acute bacterial sinusitis, acute exacerbations of chronic bronchitis, and community-acquired pneumonia. Ther Clin Risk Manag. 2006;2:59–75.

54. US Food and Drug Administration. FDA Drug Safety Communication: Important safety label changes to cholesterol-lowering statin drugs [Internet]. FDA Drug Safety Communication. 2012.

55. Dybro AM, Damkier P, Rasmussen TB, Hellfritzsch M. Statin-associated rhabdomyolysis triggered by drug-drug interaction with itraconazole. BMJ Case Rep. 2016;2016:bcr2016216457.

56. Krishna G, Ma L, Prasad P, Moton A, Martinho M, O'Mara E. Effect of posaconazole on the pharmacokinetics of simvastatin and midazolam in healthy volunteers. Expert Opin Drug Metab Toxicol. 2012;8:1–10.

57. Roques S, Lytrivi M, Rusu D, Devriendt J, De Bels D. Rhabdomyolysis-induced acute renal failure due to itraconazole and simvastatin association. Drug Metabol Drug Interact. 2011;26:79–80.

58. Rotzinger S, Baker GB. Human CYP3A4 and the metabolism of nefazodone and hydroxynefazodone by human liver microsomes and heterologously expressed enzymes. Eur Neuropsychopharmacol. 2002;12:91–100.

59. Kiser JJ, Burton JR, Anderson PL, Everson GT. Review and management of drug interactions with boceprevir and telaprevir. Hepatology. 2012;55:1620–8.

60. Konishi H, Takenaka A, Minouchi T, Yamaji A. Impairment of CYP3A4 capacity in patients receiving danazol therapy: Examination on oxidative cortisol metabolism. Horm Metab Res. 2001;33:628–30.

61. Jacobson RH, Wang P, Glueck CJ. Myositis and Rhabdomyolysis Associated With Concurrent Use of Simvastatin and Nefazodone. JAMA J Am Med Assoc. 1997;277:296–7.

62. Skrabal MZ, Stading JA, Monaghan MS. Rhabdomyolysis Associated with Simvamtatin-Nefazodone Therapy. South Med J. 2003;96:1034–5.

63. Stankovic I, Vlahovic-Stipac A, Putnikovic B, Cvetkovic Z, Neskovic AN. Concomitant administration of simvastatin and danazol associated with fatal rhabdomyolysis. Clin Ther. 2010;32:909–14.

64. Andreou ER, Ledger S. Potential drug interaction between simvastatin and danazol causing rhabdomyolysis. Can J Clin Pharmacol. 2003;10:172–4.

65. Watkins PB. The role of cytochromes P-450 in cyclosporine metabolism. J Am Acad Dermatol. 1990;23:1301–11.

66. Tseng A, Hughes CA, Wu J, Seet J, Phillips EJ. Cobicistat Versus Ritonavir: Similar Pharmacokinetic Enhancers But Some Important Differences. Ann Pharmacother. 2017;51:1008–22.

67. Scarfia RV, Clementi A, Granata A. Rhabdomyolysis and acute kidney injury secondary to interaction between simvastatin and cyclosporine. Ren Fail. 2013;35:1056–7.

68. Lasocki A, Vote B, Fassett R, Zamir E. Simvastatin-induced rhabdomyolysis following cyclosporine treatment for uveitis. Ocul Immunol Inflamm. 2007;15:345–6.

69. Chauvin B, Drouot S, Barrail-Tran A, Taburet AM. Drug-drug interactions between HMG-CoA reductase inhibitors (statins) and antiviral protease

inhibitors. Clin Pharmacokinet. 2013;52:815–31.

70. Tal A, Rajeshawari M, Isley W. Rhabdomyolysis associated with simvastatin-gemfibrozil therapy. South Med J. 1997;90:546–7.

71. Chang JT, Staffa JA, Parks M, Green L. Rhabdomyolysis with HMG-CoA reductase inhibitors and gemfibrozil combination therapy. Pharmacoepidemiol Drug Saf. 2004;13:417–26.

72. Backman JT, Kyrklund C, Kivistö KT, Wang JS, Neuvonen PJ. Plasma concentrations of active simvastatin acid are increased by gemfibrozil. Clin Pharmacol Ther. 2000;68:122–9.

73. Centers for Medicare & Medicaid Services Website. Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs).

74. Arias E, Xu J, Kochanek KD. National Vital Statistics Reports. Natl Cent Heal Stat. 2016;68.

75. AHRQ website. Agency for Healthcare Research and Quality - Healthcare Cost and Utilization Project (HCUP) [Internet]. 2019.

76. NIH CTSA website. Clinical and Translational Science Awards Program. 2016.

77. PCORI. Patient Centered Outcomes Research Institute [Internet].

78. Health Care Cost Institute [Internet].

79. OptumLabs. OptumLabs and OptumLabs Data Warehouse (OLDW) Descriptions and Citation. Cambridge, MA: n.p., May 2019. PDF. Reproduced with permission from OptumLabs.

80. Dobson AJ, Barnett AG. An Introduction to Generalized Linear Models. Chapman and Hall/CRC. 2008.

81. Boyle BH. Support vector machines: Data analysis, machine learning and applications. Nova Science Publishers, Inc. 2011.

82. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge university press. 2000.

83. Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. Vol. 18, A Journal of the Chemometrics Society. 2004. page 275–85.

84. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York: springer. 2013.

85. MacKay DJC. Bayesian Interpolation. Neural Comput. 1992;4:415–47.

86. Levenberg K. A method for the solution of certain non-linear problems in least squares. Q Appl Math. 1944;2:164–8.

87. Møller MF. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks. 1993;6:525–33.

88. Ciaburro G. Matlab for Machine Learning. Packt Publishing. 2017;

89. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. Circulation. 2017;135:e146-603.

90. Mahabaleshwarkar RK, Yang Y, Datar M V., et al. Risk of adverse cardiovascular outcomes and all-cause mortality associated with concomitant use of clopidogrel and proton pump inhibitors in elderly patients. Curr Med Res Opin. 2013;29:315–23.

91. Helin-Salmivaara A, Lavikainen PT, Korhonen MJ, et al. Pattern of statin use among 10 cohorts of new users from 1995 to 2004: A register-based nationwide study. Am J Manag Care. 2010;16:116–22.

92. Mach F, Ray KK, Wiklund O, et al. Adverse effects of statin therapy: perception vs. the evidence – focus on glucose homeostasis, cognitive, renal and hepatic function, haemorrhagic stroke and cataract. Eur Heart J. 2018;39:2526–39.

93. Bhardwaj SS, Chalasani N. Lipid-Lowering Agents That Cause Drug-Induced Hepatotoxicity. Clin Liver Dis. 2007;11:597–613.

94. Cziraky MJ, Willey VJ, McKenney JM, et al. Statin Safety: An Assessment Using an Administrative Claims Database. Am J Cardiol. 2006;97:S61–8.

95. Bellosta S, Paoletti R, Corsini A. Safety of statins: Focus on clinical pharmacokinetics and drug interactions. Vol. 109, Circulation. 2004. page III50-7.

96. Egan A, Colman E. Weighing the Benefits of High Dose Simvastatin against the Risk of Myopathy. N Engl J Med. 2011;365:285–7.

97. Josan K, McAlister FA. Cholesterol lowering for secondary prevention: What statin dose should we use? Vasc Health Risk Manag. 2007;3:615–27.

98. Floyd JS, Heckbert SR, Weiss NS, Carrell DS, Psaty BM. Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. JAMA - J Am Med Assoc. 2012;307:1580–2.

99. Andrade SE, Graham DJ, Staffa JA, et al. Health plan administrative databases can efficiently identify serious myopathy and rhabdomyolysis. J Clin Epidemiol. 2005;58:171–4.

100. Dormuth CR, Hemmelgarn BR, Paterson JM, et al. Use of high potency statins and rates of admission for acute kidney injury: Multicenter, retrospective observational analysis of administrative databases. BMJ. 2013;346:f880.

101. Layton JB, Brookhart MA, Jonsson Funk M, et al. Acute kidney injury in statin initiators. Pharmacoepidemiol Drug Saf. 2013;22:1061–70.

102. Brookhart MA, Patrick AR, Schneeweiss S, et al. Physician follow-up and provider continuity are associated with long-term medication adherence: A study of the dynamics of statin use. Arch Intern Med. 2007;167:847–52.

103. Lo Re V, Carbonari DM, Forde KA, et al. Validity of diagnostic codes and laboratory tests of liver dysfunction to identify acute liver failure events. Pharmacoepidemiol Drug Saf. 2015;24:676–83.

104. Neuvonen PJ, Niemi M, Backman JT. Drug interactions with lipid-lowering drugs: Mechanisms and clinical relevance. Clin Pharmacol Ther. 2006;80:565–81.

105. US Food and Drug Administration. Guidance for industry: Drug interaction studies study design, data analysis, implications for dosing, and labeling recommendations. US Food and Drug Administration. 2012.

106. US Food and Drug Administration. FDA Drug Safety Communication: New restrictions, contraindications, and dose limitations for Zocor (simvastatin) to reduce the risk of muscle injury [Internet]. US Food & Drug Administration. 2011.

107. Trieu J, Emmett L, Perera C, Thanakrishnan K, Van Der Wall H. Rhabdomyolysis resulting from interaction of simvastatin and clarithromycin demonstrated by Tc-99m MDP scintigraphy. Clin Nucl Med. 2004;29:803–4.

108. Lee AJ, Maddix DS. Rhabdomyolysis secondary to a drug interaction between simvastatin and clarithromycin. Ann Pharmacother. 2001;35:26–31.

109. Mesgarpour B, Gouya G, Herkner H, Reichardt B, Wolzt M. A population-based analysis of the risk of drug interaction between clarithromycin and statins for hospitalisation or death. Lipids Health Dis. 2015;14:131.

110. Graham DJ, Staffa JA, Shatin D, et al. Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. J Am Med Assoc. 2004;292:2585–90.

111. Wang YC, Hsieh TC, Chou CL, Wu JL, Fang TC. Risks of adverse events following coprescription of statins and calcium channel blockers: A nationwide population-based study. Medicine (Baltimore). 2016;95.

112. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J Chronic Dis. 1987;40:373–83.

113. Voskuijl T, Hageman M, Ring D. Higher Charlson comorbidity index scores are associated with readmission after orthopaedic surgery. Clin Orthop Relat Res. 2014;472:1638–44.

114. SooHoo NF, Farng E, Lieberman JR, Chambers L, Zingmond DS. Factors that predict short-term complication rates after total hip arthroplasty. Clin Orthop Relat Res. 2010;468:2363–71.

115. SooHoo NF, Lieberman JR, Ko CY, Zingmond DS. Factors predicting complication rates following total knee replacement. J Bone Jt Surg - Ser A. 2006;88:480–5.

116. SooHoo NF, Eagan M, Krenek L, Zingmond DS. Incidence and factors predicting pulmonary embolism and deep venous thrombosis following surgical treatment of ankle fractures. Foot Ankle Surg. 2011;17:259–62.

117. Shu CC, Lin YF, Hsu NC, Ko WJ. Risk factors for 30-day readmission in general medical patients admitted from the emergency department: A single centre study. Intern Med J. 2012;42:677–82.

118. Kurtz SM, Lau E, Ong KL, et al. Infection risk for primary and revision instrumented lumbar spine fusion in the Medicare population: Clinical article. J Neurosurg Spine. 2012;17:342–7.

119. Gabbe BJ, Magtengaard K, Hannaford AP, Cameron PA. Is the Charlson Comorbidity Index useful for predicting trauma outcomes? Acad Emerg Med. 2005;12:318–21.

120. Arrigo RT, Kalanithi P, Cheng I, et al. Charlson score is a robust predictor of 30-day complications following spinal metastasis surgery. Spine (Phila Pa 1976). 2011;36:E1274–80.

121. Kenworthy KE, Bloomer JC, Clarke SE, Houston JB. CYP3A4 drug interactions: Correlation of 10 in vitro probe substrates. Br J Clin Pharmacol. 1999;48:716–27.

122. Rodrigues a D, Roberts EM, Mulford DJ, Yao Y, Ouellet D. Oxidative Metabolism of Clarithromycin in the Presence of Human Liver Microsomes. Drug Metab Dispos. 1997;25:623–30.

123. Seithel A, Eberl S, Singer K, et al. The influence of macrolide antibiotics on the uptake of organic anions and drugs mediated by OATP1B1 and OATP1B3. Drug Metab Dispos. 2007;35:779–86.

124. Lilja JJ, Neuvonen M, Neuvonen PJ. Effects of regular consumption of grapefruit juice on the pharmacokinetics of simvastatin. Br J Clin Pharmacol. 2004;58:56–60.

125. Schmiedlin-Ren P, Edwards DJ, Fitzsimmons ME, et al. Mechanisms of enhanced oral availability of CYP3A4 substrates by grapefruit constituents: decreased enterocyte CYP3A4 concentration and mechanism-based inactivation by furanocoumarins. Drug Metab Dispos. 1997;25:1228–33.

126. Lilja JJ, Kivistö KT, Neuvonen PJ. Grapefruit juice-simvastatin interaction: Effect on serum concentrations of simvastatin, simvastatin acid, and HMG-CoA reductase inhibitors. Clin Pharmacol Ther. 1998;64:477–83.

127. Altman DG, Bland JM. Detecting skewness from summary information. BMJ. 1996;313:1200–1.

128. Rokach L, Maimon O. Decision Tree. In: Data Mining and Knowledge Discovery Handbook. 2005. page 165–92.

129. Breiman L. Random Forrests. Mach Learn. 2001;

130. Hornik K. Some new results on neural network approximation. Neural Networks. 1993;6:1069–72.

131. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks. 1989;2:359–66.

132. Smyth D, Deverall E, Balm M, Nesdale A, Rosemergy I. OUT-OF-BAG ESTIMATION. N Z Med J. 2015;128:97–100.

133. Mancini GBJ, Baker S, Bergeron J, et al. Diagnosis, Prevention, and Management of Statin Adverse Effects and Intolerance: Canadian Consensus Working Group Update (2016). Can J Cardiol. 2016;32:S35–65.

134. Castro PF, Ribeiro E, Dorea EL, Pinto GA, Hirata RDC. Factors associated with statin-related adverse muscular events in adult dyslipidemic outpatients. Brazilian J Pharm Sci. 2017;53.

135. Milionis HJ, Rizos E, Mikhailidis DP. Smoking diminishes the beneficial effect of statins: Observations from the landmark trials. Angiology. 2001;52:575–87.

Supplementary Data

Supplementary Table 1. Advantages and disadvantages of classification methods

Models	Advantages	Disadvantages
GLM	1. Response variable does not need to have a normal distribution. 2. Have more flexibility in modeling because the choice of link is separate from the probability distribution of the response variable.	1. Response variable must be independent.
SVM	1. High computational complexity for building model. Effective in high dimensional spaces. 2. It is memory efficient because of using a subset of training points (support vectors) to train model. 3. Different kernel function can be applied to fit different situations. 4. It find the global optimal of the objective function by using efficient algorithms. 5. Overfitting is addressed by maximizing the margin of the decision boundary. 6. Robust to noise.	1. SVMs do not directly provide probability estimates. 2. Difficult to handle missing values. 3. Difficult to explain.
Decision tree	1. Computationally inexpensive to construct a model and can make a quick prediction on new data even when the size of training set is large. 2. It does not make any assumptions on the data distribution. 3. Tree structure can visually represent the	1. Cannot handle missing values. 2. Overfitting is a common problem. It cannot guarantee to return the globally optimal decision tree since it only makes locally optimal decisions at each node.

	whole process of decision making which makes it easy to interpret and explain. 4. Able to handle both numerical and categorical data. 5. Can handle redundant and irrelevant features.	3. It creates biased trees if some classes dominate. Therefore, we need to balance the dataset prior to fitting with the decision tree.
Random forest	1. Runs efficiently on large databases. 2. Can effectively handle large missing data. 3. It has methods to handle unbalanced datasets. 4. The correlation between trees is avoided since a subset of the features is selected for splitting at each node. Thus, the local optimal problem of decision tree can be avoided. 5. Variance of individual tree can be reduced in random forest by taking the majority votes form all trees. 6. Can handle redundant and irrelevant features.	1. Sensitive to noisy data. Unlike decision trees, results are difficult to interpret.
ANN	1. It can handle redundant features because the weights are automatically learned during the training step and tend to be very small for redundant features. 2. Can handle irrelevant features.	1. Results may be more difficult to interpret. 2. Cannot handle missing values. 3. Sensitive to the presence of noise in the training data. 4. It is a time consuming process, especially when the number of hidden nodes is large. 5. Tends to find only locally optimum solutions.

Note: GLM = generalized linear model

SVM = support vector machine
ANN = artificial neural network

Supplementary Table 2 List of medical codes used to identify CVD events

CVD Events	ICD-9-CM Codes	HCPCS Codes
Acute myocardial infarction (AMI)	410.01, 410.11, 410.21, 410.31, 410.41, 410.51, 410.61, 410.71, 410.81, 410.91	
Unstable Angina	411.81	
Coronary artery disease	414.0, 414.01, 414.02, 414.04, 414.05, 414.2, 414.3, 414.4, 414.8	
Stroke	433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, 434.91, 435.8, 435.9	
Percutaneous Coronary Intervention (PCI)	00.66, 36.06, 36.07	92973, 92974, 92975, 92978, 92979, 92980, 92981, 92982, 92984, 92995, 92996, G0290, G0291
Coronary artery bypass graft surgery (CABG)	36.1, 36.10, 36.11, 36.12, 36.13, 36.14, 36.15, 36.16, 36.17, 36.19, 36.2	33510, 33511, 33512, 33513, 33514, 33516, 33517, 33518, 33519, 33521, 33522, 33523, 33533, 33534, 33535, 33536, S2205, S2206, S2207, S2208, S2209

Supplementary Table 3 74 Extracted features

Category	Number of features	Examples
Patient information	7	Age, gender, home ownership, health insurance type[1], region, Charlson comorbidity score, Elixhauser comorbidity score
Prescription information	10	Atorvastatin drug strength, fluvastatin drug strength, lovastatin drug strength, pitavastatin drug strength, pravastatin drug strength, rosuvastatin drug strength, simvastatin drug strength, statin exposure time, formulary drug, mail order
Cost	3	Medical costs (first 30 days) [2], health plan paid statin costs (per 30 days), statin out-of-pocket cost (per 30 days)
Others	7	Initiation gap[3], prescriber specialty, transferred claim[4], preferred benefits[5], capitation status[6], inpatient CVD claim[7], outpatient or office visit CVD claim[8]
Comorbidities	47	17 Charlson comorbidities (see supplementary table 6), 30 Elixhauser comorbidities (see supplementary table 7)

Note:
1. Health insurance type: commercial or Medicare Advantage;
2. Medical cost (first 30 days): total cost of health plan paid and patient out-of-pocket cost for the first 30 days since CVD index date;
3. Initiation gap is the gap between statin index date and CVD index date;
4. Transferred claim: discharge status. Discharged and transferred such as home, nursing facility, and other places;
5. Preferred benefits: preferred level of reimbursement on the claim;
6. Capitation status: identifies if the service is fee-for-service or capitated.
7. Inpatient CVD claims: CVD inpatient claim(s) on CVD index date or the following 2 days;
8. Outpatient or office visit CVD claim: Only had outpatient or office visit CVD claims on CVD index date without any CVD claims (inpatient/outpatient/office visits) on CVD index date and the following 90 days.

Supplementary Table 4. Performance of single features

Order	Features
1-10	Age, Charlson comorbidity score, initiation gap, statin exposure time, health insurance type, Elixhauser comorbidity score, statin out-of-pocket cost, medical cost, gender, Hypertension
11-20	inpatient CVD claim, preferred benefits, prescriber specialty, chronic pulmonary disease (Charlson), fluid and electrolyte disorders, chronic pulmonary disease, deficiency anemia, renal disease (Charlson), diabetes (uncomplicated), Diabetes without chronic complication (Charlson)
21-30	Renal failure, other neurological disorders, peripheral vascular disease (Charlson), hypothyroidism, peripheral vascular disorders, outpatient or office visit CVD claim, region, congestive heart failure (Charlson), depression, congestive heart failure
31-40	Diabetes (complicated), cerebrovascular disease (Charlson), cardiac arrhythmias, rheumatoid arthritis/collagen vascular diseases, diabetes with chronic complication (Charlson), drug formulary, simvastatin drug strength, cancer (Charlson), mild liver disease (Charlson), Psychoses
41-50	Solid tumor without metastasis, rosuvastatin drug strength, rheumatic disease (Charlson), weight loss, valvular disease, pulmonary circulation disorders, obesity, mail order, coagulopathy, dementia (Charlson),
51-60	Liver disease, drug abuse, blood loss anemia, paralysis, lymphoma, atorvastatin drug strength, transferred claim, lovastatin drug strength, hemiplegia or paraplegia (Charlson), metastatic cancer
61-70	Metastatic solid tumor (Charlson), myocardial infarction (Charlson), home ownership, moderate or severe liver disease (Charlson), alcohol abuse, peptic ulcer disease (Charlson), fluvastatin drug strength, pitavastatin drug strength, capitation, peptic ulcer disease excluding bleeding,
71-74	AIDS/HIV, AIDS/HIV (Charlson), health plan paid statin cost, pravastatin drug strength

Supplementary Table 5. Features selected by forward selection and clinical expert

	Forward Feature Selection (N=35)	Expert selection (N=32)
Age	√	√
Gender	√	√
Charlson comorbidity score	√	√
Statin exposure time	√	√
Health insurance type	√	√
Medical cost	√	√
Statin initiation gap	√	√
Prescriber specialty	√	√
Region	√	
Mail order	√	
Drug formulary	√	√
Statin out-of-pocket cost	√	√
Simvastatin drug strength	√	√
Lovastatin drug strength	√	√
Fluvastatin drug strength	√	√
Atorvastatin drug strength		√
Pitavastatin drug strength		√
Pravastatin drug strength		√
Rosuvastatin drug strength		
Preferred benefits	√	√
Inpatient CVD claim	√	√
Transferred claim	√	
Home ownership		√
Capitation status		√
Fluid and electrolyte disorders	√	√
Rheumatoid arthritis/collagen vascular diseases	√	
Metastatic cancer	√	
Other neurological disorders	√	
Renal failure	√	√
Psychoses	√	√
Deficiency anemia	√	√
Cerebrovascular disease	√	
Hypothyroidism	√	
Weight loss	√	√
Paralysis	√	
Lymphoma	√	
Depression	√	√
Solid tumor without	√	

metastasis		
Alcohol abuse		√
Congestive heart failure		√
Coagulopathy		√
Drug abuse		√
Liver disease		√
Pulmonary circulation disease		√
Mild liver disease (Charlson)	√	
Cancer (Charlson)	√	
Diabetes with chronic complication (Charlson)	√	

Supplementary Table 6. ICD-9-CM coding algorithms and weights for Charlson comorbidities

Comorbidities	ICD-9-CM	Weights
Myocardial infarction	410.x, 412.x	1
Congestive heart failure	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4-425.9, 428.x	1
Peripheral vascular disease	093.0, 437.3, 440.x, 441.x, 443.1-443.9, 447.1, 557.1, 557.9, V43.4	1
Cerebrovascular disease	362.34, 430.x-438.x	1
Dementia	290.x, 294.1, 331.2	1
Chronic pulmonary disease	416.8, 416.9, 490.x-505.x, 506.4, 508.1, 508.8	1
Rheumatic disease	446.5, 710.0-710.4, 714.0-714.2, 714.8, 725.x	1
Peptic ulcer disease	531.x-534.x	1
Mild liver disease	070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.6, 070.9, 570.x, 571.x, 573.3, 573.4, 573.8, 573.9, V42.7	1
Diabetes without chronic complication	250.0-250.3, 250.8, 250.9	1
Diabetes with chronic complication	250.4-250.7	2
Hemiplegia or paraplegia	334.1, 342.x, 343.x, 344.0-344.6, 344.9	2
Renal disease	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 582.x, 583.0-583.7, 585.x, 586.x, 588.0, V42.0, V45.1, V56.x	2
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	140.x-172.x, 174.x-195.8, 200.x-208.x, 238.6	2

Moderate or severe liver disease	456.0-456.2, 572.2-572.8	3
Metastatic solid tumor	196.x-199.x	6
AIDS/HIV	042.x-044.x	6

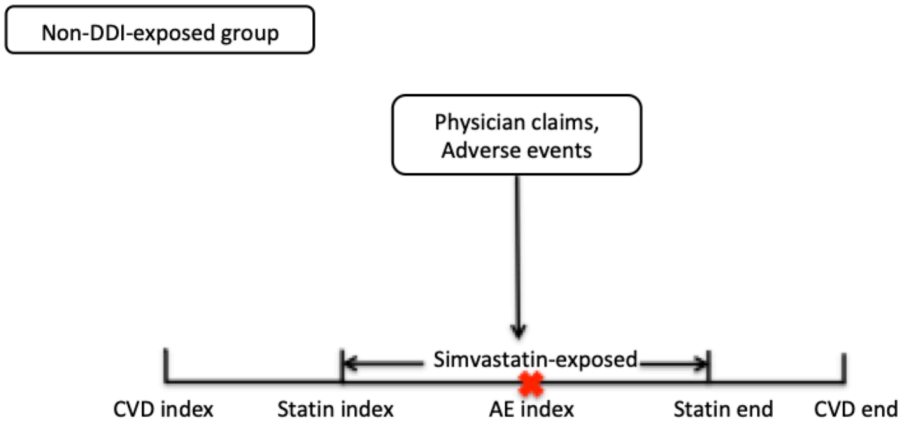
Note: Comorbidities were identified based on occurrence of at least one inpatient diagnose or two outpatient diagnoses. Each comorbidity category is assigned a weight (1, 2, 3, or 6) according to their potential influence on mortality. The sum of all the weights results in a single comorbidity score - the Charlson comorbidity score for a patient to predict the outcome and risk of death from many comorbid diseases. A score of zero indicates that no comorbidities were found. The higher the score, the more likely the predicted outcome will result in mortality.

Supplementary Table 7. ICD-9-CM coding algorithms and weights for Elixhauser comorbidities

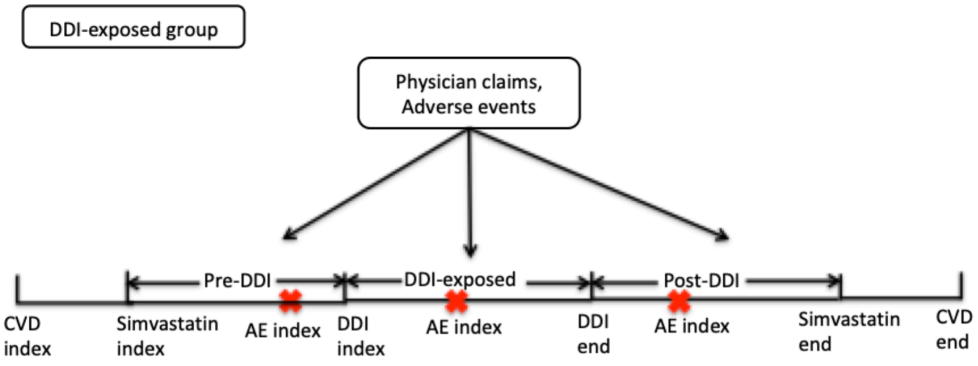
Comorbidities	ICD-9-CM	Weights
Congestive heart failure	398.91, 402.11, 402.91, 404.11, 404.13, 404.91, 404.93, 428.x	7
Cardiac arrhythmias	426.10, 426.11, 426.13, 426.2-426.53, 426.6-426.8, 427.0, 427.2, 427.31, 427.60, 427.9, 785.0, V45.0, V53.3	5
Valvular disease	093.2, 394.0-397.1, 424.0-424.91, 746.3-746.6, V42.2, V43.3	-1
Pulmonary circulation disorders	416.x, 417.9	4
Peripheral vascular disorders	440.x, 441.2, 441.4, 441.7, 441.9, 443.1-443.9, 447.1, 557.1, 557.9, V43.4	2
Hypertension,	401.1, 401.9, 402.10, 402.90, 404.10, 404.90, 405.1, 405.9	0
Paralysis	342.0, 342.1, 342.9-344.x	7
Other neurological disorders	331.9, 332.0, 333.4, 333.5, 334.x, 335.x, 340.x, 341.1-341.9, 345.0, 345.1, 345.4, 345.5, 345.8, 345.9, 348.1, 348.3, 780.3, 784.3	6
Chronic pulmonary disease	490-492.8, 493.00-493.91, 494.x-505.x, 506.4	3
Diabetes, uncomplicated	250.0-250.3	0
Diabetes, complicated	250.4-250.7, 250.9	0
Hypothyroidism	243-244.2, 244.8, 244.9	0
Renal failure	403.11, 403.91, 404.12, 404.92, 585.x, 586.x, V42.0, V45.1, V56.0, V56.8	5
Liver disease	070.32, 070.33, 070.54, 456.0, 456.1, 456.2, 571.0, 571.2-571.9, 572.3, 572.8, V42.7	11
Peptic ulcer disease excluding bleeding	531.70, 531.90, 532.70, 532.90, 533.70, 533.90, 534.70, 534.90, V12.71	0
AIDS/HIV	042.x–044.x	0
Lymphoma	200.x–202.3x, 202.5–203.0, 203.8, 238.6, 273.3, V10.71, V10.72, V10.79	9
Metastatic cancer	196.x–199.x	12
Solid tumor without metastasis	140.x–172.x, 174.x, 175.x, 179.x–195.x, V10.x	4
Rheumatoid	701.0, 710.x, 714.x, 720.x, 725.x	0

arthritis/collagen vascular diseases		
Coagulopathy	286.x, 287.1, 287.3–287.5	3
Obesity	278.0	4
Weight loss	260.x–263.x	6
Fluid and electrolyte disorders	276.x	5
Blood loss anemia	280.0	-2
Deficiency anemia	280.1–281.9, 285.9	-2
Alcohol abuse	291.1, 291.2, 291.5–291.9, 303.9, 305.0, V113	0
Drug abuse	292.0, 292.82–292.89, 292.9, 304.0, 305.2–305.9	7
Psychoses	295.x–298.x, 299.1	0
Depression	300.4, 301.12, 309.0, 309.1, 311	3

Note: Comorbidities were identified based on occurrence of at least one inpatient diagnose or two outpatient diagnoses. Each comorbidity category is assigned a weight from -2 to 12 according to their potential influence on mortality. The sum of all the weights results in a single comorbidity score - the Elixhauser comorbidity score for a patient to predict the outcome and risk of death from many comorbid diseases. The higher the score, the more likely the predicted outcome will result in mortality.



Note:
1. Physician claims are identified and counted simvastatin-exposed period.
2. Index adverse event identified during simvastatin-exposed period.



Note:
1. Physician claims are counted during pre-DDI, DDI-exposed, and post-DDI time periods.
2. Index adverse events are identified during pre-DDI, DDI-exposed, and post-DDI time periods.
3. DDI unexposed period includes pre-DDI and post-DDI period.

Supplementary Figure 1. Demonstration of key dates, index dates, and time periods.